The Association for Literary and Linguistic Computing
The Association for Computers and the Humanities
The Society for Digital Humanities — Société pour l'étude des médias interactifs

Digital Humanities 2009

Conference Abstracts

University of Maryland, College Park
June 22 – 25, 2009

The 21st Joint International Conference of the Association for Literary and Linguistic Computing, and the Association for Computers and Humanities

and

The 2nd Joint International Conference of the Association for Literary and Linguistic Computing, the Association for Computers and Humanities, and the Society for Digital Humanities — Société pour l'étude des médias interactifs

International Program Committee

- Clair Warwick (ACH, chair)
- Brett Barney (ACH)
- Michael Eberle-Sinatra (SDH-SEMI)
- Willard McCarty (ACH)
- John Nerbonne (ALLC)
- Allen Renear (ACH)
- Jan Rybicki (ALLC)
- Stéfan Sinclair (SDH-SEMI)
- Paul Spence (ALLC)

Local Organizers

- Professor Neil Fraistat (Director, MITH and English)
- Associate Professor Matthew Kirschenbaum (Associate Director, MITH and English)
- Kate Singer, Conference Coordinator (MITH and English)

ISBN: 978-061-52-9929-7
Published by the Maryland Institute for Technology in the Humanities (MITH)

Cover design: Greg Lord, MITH
Editorial & Layout: Kate Singer, MITH

The Paradox of the Programme Committee

The job of the programme committee this year has been a paradoxical one. We were delighted to receive such a large number of proposals of such high quality on such a wide range of topics. It was especially pleasing to see how many submissions we received in the area of authorship studies and stylistics. As a result, papers in this area constitute a strong theme in this year's conference.

The quality of the submissions made our job a pleasant one, since we were delighted to see the evidence of a discipline that is flourishing across the world. However, we had many more very good submissions than we had space for in the programme, even though the local organisers kindly agreed to add more sessions. Despite protracted discussion and agonising over the last few slots, we had to turn down several potentially exciting sessions. Although this was the less pleasant aspect of the committee's job, it is also evidence of the very high standard of research and application development in the field. So we hope that our pain is your gain, since the sessions we did choose promise to be truly excellent: challenging, stimulating and exciting for a whole range of Digital Humanists.

It will be evident that my heartfelt thanks are due to all the Programme Committee members for their hard work this year, and especially to the vice chair John Nerbonne. I am also hugely grateful to Sara Schmidt for all her help with the conftool system, and for being a one woman repository of information about all aspects of DH programme planning. The local organisers, Neil Fraistat and Matt Kirschenbaum, Kate Singer and their team have been a model of flexibility, responsiveness and helpfulness, so thanks to all of you.

I hope this will be a truly memorable conference and that you will find all aspects of it thoroughly enjoyable. Most importantly I would like to thank in advance all the presenters, without whom none of this would have been possible.

Thank you and welcome to the conference!

Claire Warwick
Chair of the International Programme Committee

A Letter from the Local Organizers

Dear DHers,

On behalf of the Maryland Institute for Technology in the Humanities and the entire University of Maryland community, we welcome you to Digital Humanities 2009.

Hosting the annual conference of the Alliance of Digital Humanities Organizations would be a privilege under any circumstances. In our case, it bears special significance because 2009 also marks MITH's 10th year as a working digital humanities center. We are thrilled to be able to celebrate this anniversary with nearly 300 of our closest digital humanities friends!

The papers, panels, and posters you will be enjoying over the next several days were selected by an international program committee which has overseen an extremely competitive submission process to produce a conference of the highest intellectual quality. We are honored to furnish the venue for the work about to be presented. Our two keynote speakers, Lev Manovich and Christine Borgman, will serve as touchstones for discussion and debate while also reflecting signature elements in digital humanities here at Maryland, namely new media studies and our close ties with the College of Information Studies (the iSchool). We are also pleased to be able to welcome a number of exhibitors to the conference.

We are especially looking forward to these printed proceedings acquiring their dynamic and spontaneous mirror life online. In order to lend some coherence to the content that will be generated by many of you, and to enable others elsewhere to better follow events, please use dh09 as the official conference tag for all your posts, tweets, photos, videos, and whatever other media and messages you contribute to the Web 2.0 cloud. You may also wish to follow the official conference Twitter feed, @dh09.

While Digital Humanities 2009 will not take place within the confines of MITH itself, we wanted to say a few words about the space that has put Maryland on the global digital humanities map. A collaboration among the University of Maryland's College of Arts and Humanities, Libraries, and Office of Information Technology, since its founding in 1999 MITH has become internationally recognized as one of the leading centers of its kind, distinguished by the cultural diversity so central to its identity. Located in McKeldin Library at the heart of the campus, MITH is the University's primary intellectual hub for scholars and practitioners of digital humanities, electronic literature, and cyberculture, as well as the home of the Electronic Literature Organization, the most prominent international group devoted to the writing, publishing and reading of electronic literature.

MITH faculty, fellows, and staff have served as principle investigators, co-principle investigators, and sub-contracts on grants and awards from the NEH, the IMLS, the NSF, the Andrew W. Mellon Foundation, and the Library of Congress, among others. With projects that range (literally) from Shakespeare to Second Life, our partners include the Folger Shakespeare Library, the Bodleian Library, the British Library, the University of Illinois Urbana-Champaign, Stanford University, Rice University, the Harry Ransom Center, Emory University, George Mason's Center for History and New Media, and Linden Lab. MITH's research and intellectual mission is complemented by its public programs and events. In addition to our popular Digital Dialogues series—which has featured nearly 100 speakers to date—we have hosted or co-hosted such recent events as the Future of Electronic Literature, Digital Diasporas: Digital Humanities and African American/African Diaspora Studies, a Summit of Directors of Digital Humanities Centers (with NEH), and Tools for Data-Driven Scholarship (with CHNM at George Mason).

Digital Humanities 2009 continues this tradition, but it would not have been possible without the labor and expertise of some key individuals. The international program committee chaired by Claire Warwick has been tireless in its efforts on behalf of the conference, and a uniform pleasure to work with. Kate Singer, who earned her Ph.D. from Maryland in May, has overseen every detail of our planning and preparation. Seemingly no detail has ever been further away than her fingertips. Greg Lord is the guiding hand and imaginative eye behind all of the visual design and imagery for the conference. Chris Grogan and Doug Reside at MITH have also provided invaluable assistance and support. Our spirited student volunteers have given their time and sweat of the brow. Lisa Lena Opas-Hänninen and John Unsworth have shared their wisdom and experience from their recent conferences, as has Sara Schmidt from Illinois. Allison Druin, Jenny Preece, and Martha Nell Smith graciously served on our local advisory committee. Last but not least, Lisa Press and Alison Nagle at Conferences and Visitor Services have kept the foundation in place and the roof from coming off. Our generous sponsors we are pleased to acknowledge elsewhere in these pages.

Best wishes for a productive, rewarding, and exciting conference here in College Park!

Sincerely,

Neil Fraistat and Matthew Kirschenbaum
Co-Local Organizers, Digital Humanities 2009

The Local Organizing Committee
would like to acknowledge the following institutions
for their generous support

University of Maryland, Division of Research

University of Maryland, College of Arts and Humanities

University of Maryland Libraries

University of Maryland, Department of English

University of Maryland, College of Information Studies

University of Maryland, Human-Computer Interaction Lab

Allied Digital Humanitites Organization (ADHO)

Oxford University Press

Table of Contents

Plenary Sessions

Panels

Papers

Posters

Reviewers

Akama, Hiroyuki
Anderson, Dr. Ian
Anderson, Jean Gilmour
Andreev, Vadim Sergeevich
Baayen, Prof. Rolf Harald
Barney, Brett
Bauman, Syd
Baumann, Ryan Frederick
Bearman, David
Beavan, David
Bentkowska-Kafel, Dr. Anna
Biber, Dr. Hanno
Birnbaum, Prof. David J
Blanke, Dr. Tobias
Bodard, Dr. Gabriel
Bodenhamer, Dr. David
Bol, Prof. Peter Kees
Bolter, Prof. Jay David
Boot, Peter
Bosse, Arno
Bowen, Prof. William
Bradley, John
Brey, Gerhard
Brown, Prof. Susan
Buchmüller, Sandra
Burnard, Lou
Burr, Prof. Elisabeth
Bush, Chuck
Cantara Abbott, Linda
Caton, Dr. Paul
Cayless, Dr. Hugh
Ciula, Dr. Arianna
Clement, Tanya
Conner, Prof. Patrick
Connors, Louisa
Cooney, Dr. Charles M.
Cooper, Dr. David Christopher
Cossard, Prof. Patricia Kosco
Craig, Prof. Hugh
Cummings, Dr. James C.
Dahlstrom, Dr. Mats
David, Stefano
Dawson, Dr. John
Devlin, Dr. Kate
DiNunzio, Joseph
Dombrowski, Quinn Anya
Downie, Prof. J. Stephen
Dubin, Dr. David S.
Dunn, Dr. Stuart
Durand, Dr. David G.
Durusau, Patrick
Eberle-Sinatra, Dr. Michael
Edmond, Dr. Jennifer C
Egan, Dr. Gabriel
Eide, Øyvind
Ell, Dr. Paul S
Esteva, Maria
Fischer, Dr. Franz
Flanders, Dr. Julia
Forest, Prof. Dominic
Furuta, Dr. Richard
Galina Russell, Isabel
Gallet-Blanchard, Prof. Liliane
Gants, Prof. David
Gartner, Richard
Giordano, Dr. Richard
Goldfield, Dr. Joel
Gow, Ann
Groß, Dr. Nathalie
Gueguen, Gretchen Mary
Hawkins, Kevin Scott
Hockey, Prof. Susan
Holmes, Martin
Hoover, Dr. David L.
Hughes, Lorna
Hunyadi, Prof. László
Hyman, Dr. Malcolm D.
Isaksen, Leif
Ivanovs, Dr. Aleksandrs
Jessop, Martyn
Jockers, Dr. Matthew
Johnsen, Dr. Lars
Johnson, Dr. Ian R.
Juola, Prof. Patrick
Kansa, Prof. Eric Christopher
Kansa, Dr. Sarah Whitcher
Keating, Dr. John Gerard
Kretzschmar, Dr. William
Lancaster, Dr. Lewis Rosser
Lavagnino, Dr. John
Lavrentiev, Dr. Alexei
Leitch, Caroline
Litta Modignani Picozzi, Dr. Eleonora
Luyckx, Kim
Lüngen, Dr. Harald
Mahony, Simon
Martin, Prof. Worthy N.
Martinet, Prof. Marie-Madeleine
McCarty, Prof. Willard
McDaniel, Dr. Rudy
Meister, Prof. Jan Christoph
MendezRodriquez, Dr. Eva
Meschini, Federico
Miles, Adrian

Mostern, Dr. Ruth
Mylonas, Elli
Nagasaki, Kiyonori
Nerbonne, Prof. John
Newton, Greg T.
O'Donnell, Dr. Daniel Paul
Olsen, Prof. Mark
Opas-Hänninen, Prof. Lisa Lena
Ore, Espen S.
Pasanek, Brad
Pidd, Michael
Pierazzo, Dr. Elena
Piez, Dr. Wendell
Porter, Dorothy Carr
Pytlik Zillig, Prof. Brian L.
Rahtz, Sebastian
Rains, Michael John
Ramsay, Dr. Stephen
Rehbein, Malte
Remnek, Prof. Miranda
Renear, Dr. Allen H.
Reside, Dr. Doug
Robinson, Prof. Peter
Rockwell, Prof. Geoffrey
Roe, Glenn H
Romary, Prof. Laurent
Rudman, Prof. Joseph
Ruecker, Dr. Stan
Ruotolo, Christine
Rybicki, Dr. Jan
Schmidt, Harry
Schreibman, Prof. Susan
Sculley, D.
Shawver, Dr. Gary
Siemens, Dr. Raymond George
Simons, Prof. Gary F.
Sinclair, Prof. Stéfan
Smith, David A.
Smith, Prof. Martha Nell
Smith, Natalia (Natasha)
Snyder, Dr. Lisa M.
Spence, Paul Joseph
Sperberg-McQueen, Dr. Michael
Spiro, Dr. Lisa
Stauffer, Stephanie J.
Steggle, Prof. Matthew
Stokes, Dr. Peter Anthony
Sukovic, Suzana
Suzuki, Takafumi
Terras, Dr. Melissa
Thaller, Prof. Manfred
Tripp, Mary L.
Unsworth, John
Van den Branden, Ron
Van Elsacker, Bert
Vanhoutte, Edward
Váradi, Dr. Tamás
Walsh, Prof. John
Warwick, Dr. Claire
Wiesner, Dr. Susan L.
Wilkens, Matthew
Willett, Perry
Winder, Dr. William
Witt, Dr. Andreas
Wolff, Prof. Mark
Worthey, Glen
Yu, Dr. Bei
Zafrin, Dr. Vika
Zhang, Junte
Zimmerman, Matthew

Keynote Address

Monday, June 22, 2009

Activating The Archive, Or: Data Dandy Meets Data Mining

Lev Manovich
Professor, Visual Arts Department, University of California, San Diego
http://www.manovich.net/

Abstract

The joint availability of massive amounts of digitized cultural heritage as well as all the born-digital content (along with the data about people's production, sharing, and reception of this content) allows for the new ways of researching, teaching, and exhibiting culture. What are the theoretical and methodological issues that rise when we start treating culture as data that can be automatically analyzed and visualized? What are the consequences of treating a pattern as a new basic epistemological element of knowledge? Is it sufficient to borrow the techniques from the fields of computer science, information visualization and media art – or do we need to develop new techniques specific to humanities? How do we address the new "data divide"- between the people and cultural processes which leave rich digital traces (and therefore will be analyzed and written about) and those which do not?

In my talk I will address these and other conceptual issues around "cultural data mining." My focus will be on two emerging areas: analysis of visual media and analysis of born digital and web native content. I will demonstrate the techniques developed at Software Studies Initiative at Calit2/UCSD for the analysis and visualization of patterns in images and video - feature films, cartoons, television art, user-generated video, etc. I will discuss challenges and new exiting possibilities which arise when we start looking at web sites and blogs, social media sites, digital art, games and other interactive media. I will also show the results emerging from our large scale study of cinema, video games and social media which we are currently undertaking at NERSC (National Department of Energy Supercomputer Center) with the support from NEH Humanities High Performance Computing grant.

Biography

Lev Manovich's books include *Software Takes Command* (released under CC license, 2008; forthcoming from The MIT Press), *Soft Cinema: Navigating the Database* (The MIT Press, 2005), and *The Language of New Media* (The MIT Press, 2001) which is hailed as "the most suggestive and broad ranging media history since Marshall McLuhan." He has written 100 articles which have been reprinted over 300 times in 30+ countries. Manovich is a Professor in Visual Arts Department, University of California-San Diego, a Director of the Software Studies Initiative at California Institute for Telecommunications and Information Technology (Calit2), and a Visiting Research Professor at Godsmith College (University of London), De Montfort University (UK), and University of New South Wales (Sydney).

Keynote Address
Tuesday, June 23, 2009

Scholarship in the Digital Age: Blurring the Boundaries between the Sciences and the Arts

Christine L. Borgman
Professor & Presidential Chair in Information Studies
University of California, Los Angeles
http://is.gseis.ucla.edu/cborgman

Abstract

As the digital humanities mature, their scholarship is taking on many characteristics of the sciences, becoming more data-intensive, information-intensive, distributed, multi-disciplinary, and collaborative. While few scholars in the humanities or arts would wish to be characterized as emulating scientists, they do envy the comparatively rich technical and resource infrastructure of the sciences. The interests of all scholars in the university align with respect to access to data, library resources, and computing infrastructure. However, the scholarly interests of the sciences and humanities diverge regarding research practices, sources of evidence, and degrees of control over those sources. This talk will explore the common and competing interests of disciplines for scholarship in the digital age.

Biography

Christine L. Borgman is Professor and Presidential Chair in Information Studies at UCLA. She is the author of more than 180 publications in the fields of information studies, computer science, and communication. Both of her sole-authored monographs, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (MIT Press, 2007) and *From Gutenberg to the Global Information Infrastructure: Access to Information in a Networked World* (MIT Press, 2000), have won the Best Information Science book of the year award from the American Society for Information Science and Technology. She is a lead investigator for the Center for Embedded Networked Systems (CENS), a National Science Foundation Science and Technology Center, where she conducts data practices research. She chaired the Task Force on Cyberlearning for the NSF, whose report, *Fostering Learning in the Networked World*, was released in July, 2008. Prof. Borgman is a fellow of the American Association for the Advancement of Science and a member of the US National Academies' Board on Research Data and Information.

Panels

Digital Classicist: Re-use of Open Source and Open Access Publications in Ancient Studies

Chair: Gabriel Bodard
King's College London

Those of us who publish humanities data in digital form often make the claim that one of many advantages of electronic publication is the ability to make available source data for re-use and analysis by future scholars. If the source data, and possibly also the tooling or a processing statement, is made not only available but licensed for re-use, this potentially allows asynchronous collaborators, reviewers, and others to test the published conclusions, to apply different assumptions to the data. Where the digital source and processes are an essential part of the commentary published and conclusions drawn, it is arguable that it would be academically irresponsible not to make these resources available for replication and testing of ones conclusions. Humanists have always recognise the importance of publishing with full bibliography, history of scholarship, and critical apparatus; but we might also learn from the physical sciences where experimental methodology and raw data are essential elements of the publication of any research.

This is a solid general statement, and I have made these sorts of arguments myself (e.g. in DM 4 [2008]; Bodard/Garcés in M. Deegan & K. Sutherland, *Text editing, print, and the digital world* [2009]), but there is often relatively little evidentiary support in the form of openly published datasets that have been independently tested or re-used by other projects. In this panel we aim to bring together several examples of the re-use of datasets relating to the ancient world by projects other than those that created them. The participants in this panel have all either (a) published data or developed tools under an Open Access/Source license, or (b) made use of Open Access/Source materials in original research projects of their own.

- The issues we shall address in this panel will include:
- The importance of open licensing in addition to merely making material "free" (not having to ask permission);
- Electronic publication as resource creation *versus* self-contained research output;
- Advantages of publishing source code and methodology as well as polished output of data and conclusions;
- Enabling re-uses that cannot be predicted by the creator of the original product;
- Re-use strategies: improving access or interface *versus* creating new interpretations or aggregations;
- Re-use as non-concurrent collaboration, improving data and interpretation;
- Issues of re-publication: attribution, versioning, and forking.

The papers in this panel stem from very different projects with a range of approaches and agendas. The LaQuAT project is based almost entirely upon re-use of published data, and so relies on the Open Access publication of primary sources (or the goodwill of scholars where data is incomplete or unlicensed). Pleiades is creating data, or newly aggregating it from multiple scholarly sources, to publish under a Creative Commons license (Attribution-ShareAlike), and exploit and contribute to several Open Source software projects. The Homer Multitext is a project that both relies on open standards and tools, and produces large quantities of open-licensed raw data. All are projects that value collaboration, both direct in terms of working with colleagues in the same or other disciplines, and indirect in the sense of producing scholarly outputs that are conducive to building upon, adapting testing, and re-using.

Paper 1: Linking and Querying Ancient Texts: a case study with three epigraphic/papyrological datasets

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

Tobias Blanke
King's College London
tobias.blanke@kcl.ac.uk

Mark Hedges
King's College London
mark.hedges@kcl.ac.uk

The OGSA-DAI (Open Grid Service Architecture—Data Access and Integration, http://www.ogsadai.org.uk/) project supports the exposure of data resources, such as relational or XML databases, on to grids. Various interfaces are provided and many database management systems are supported, with a particular view to querying, transforming and delivering data in different ways via a simple toolkit for developing client applications. OGSA-DAI is designed to be extensible, so users can provide their own additional functionality.

Colleagues at the Edinburgh Parallel Computing Centre and the Centre for e-Research at KCL have been funded to carry out a small case study applying the OGSA-DAI platform to three datasets of ancient texts in different formats. The Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV) is a collection of metadata (largely bibliographic, geographical, and dating) for 65000 Greek papyri from Egypt, stored in a large Filemaker Pro database. The Project Volterra is a database of legal texts from the Roman empire, currently in the low tens of thousands but very much in progress, stored in a series of themed tables in MS Access. The Inscriptions of Aphrodisias (IAph) is a corpus of just under 2000 ancient Greek inscriptions from a single city in Asia Minor, published in TEI XML. These collections span roughly the same period - the first five centuries or so of the Roman Empire - and also overlap in terms of places and people, although their contents are otherwise quite different. The provision of an integrated view would thus be fruitful for the researcher. A particularly challenging issue being investigated is that of handling different levels of uncertainty in temporal data: some dates are extremely precise – even to the day – whereas many others are very vague – perhaps to a span of 50 or 100 years.

These datasets are all freely available in one form or another, and the scholars who own the databases are happy for us to re-use them in this way and publish the results of our aggregation and federated querying. In an ideal world, of course, we should not have to seek permission from the owners at all in order to re-use and re-purpose their published data. The IAph texts are all published under a Creative Commons-Attribution licence (CC-BY), so re-use is not only permitted but encouraged (in fact Bodard is one of the authors of this dataset, but in any case we can use these texts for anything we like without asking or even informing the authors so long as we attribute the original material to the copyright holders). A transformation of the HGV data into EpiDoc XML has likewise been published under CC-BY, although it is the master database that interests us for this project, and that is not publicly available in its raw form (although a HTML version is online and free). There is also a free, web-available version of the Volterra data (although the website is down at time of writing), but the database itself was acquired for this project with the permission of the editors.

As mentioned above the contents of these three datasets vary quite widely, but there is sufficient overlap to enable a certain amount of cross-database searching to be feasible, at least as a proof of principle. For instance, although the Volterra database specifically addresses legal texts, it contains some papyri and thus possibly references to places that also occur in the HGV metadata. Likewise, although the Volterra texts do not include any inscriptions from Aphrodisias, there may be attestations of persons that appear in both the Volterra and IAph texts (especially in the late antique period, which is where the Aphrodisias material is most richly annotated). IAph and HGV do not directly share any content, but the categories that are used to organize the texts have a certain overlap, for example letters, decrees, honours, contracts. As mentioned above, all three datasets overlap fairly closely in date, and have similar (but not identical) mechanisms for recording dates, date-ranges, periods, and uncertain dating. Cross-corpus search in all of these areas or combinations of them should test the OGSA-DAI software and demonstrate the validity and usefulness of this approach.

OGSA-DAI is considered to be a standard for database integration in Grid environments, which enable virtualisation and sharing of resources via the Internet, as well as in a purely web-service environment. Until now, the OGSA-DAI technology has been used mainly to provide integrated views of relational databases with different schemas, and the LaQuAT demonstrator will to begin with use it in this way with the two database resources, HGV and Projet Volterra. Subsequently the work will be extended to integrate the InsAph XML files, providing an integrated view over the three three structured data resources. The project will also produce significant enhancements to the OGSA-DAI software, specifically in its handling of XML resources, which is currently more restricted than its features for database integration. OGSA-DAI will then integration of multiple database and XML data resources. The LaQUaT demonstrator will use a recent extension to OGSA-DAI called OGSA-DQP, which is a service-based distributed query processor, to produce queries across these data resources.

One project output will be thus be an openly available demonstrator allowing an integrated view over these three datasets. However, the resources selected are just examples from among numerous others to which the LaQuAT approach could be applied. In the fields of archaeology and classics alone, there are numerous datasets, often small and isolated, that would be of great utility if the information they contained could be integrated. Three points to note about many of these resources are that:

- Formats are very diverse. The databases rarely follow standardised database schemas, so typically any two schemas will be different. Moreover, use of mark-up can vary significantly, particularly in older resources before much effort had been towards standardisation (such as EpiDoc), but stylistic variation may occurs even when standards are applied.

- Resources are not easily available for use; they may locked away on local or departmental machines, or "published" on a website in a way that is not particularly usable by a researcher.

- Even when a resource is available it is often available only in isolation. Many of these resources may be regarded as fragments of a larger picture, with vastly more value if researchers could access this larger picture rather than just the parts.

- Resources may be owned by different communities and subject to different rights; the scholars who created them may be unwilling to accept anything that affects the integrity of the original resources. Consequently, any integration initiative must respect this autonomy and integrity, if it is to be successful.

The ability to link up such diverse data resources, in a way that respects the original data resources and the communities responsible for them, is a pressing need among humanities researchers. The LaQuAT project is developing a software demonstrator utilising a small set of resources in a particular discipline; however, the solution developed will have a lifespan beyond the initial project and will provide a framework into which other researchers will be able to attach resources of interest, thus building up a critical mass of related material whose utility as a research tool will be significantly greater than that of the sum of its parts. We see this project as providing an opportunity to start building a more extensive e-infrastructure for advanced research in the (digital) humanities. Once humanities scholars are persuaded of the feasibility of this approach, there are many other datasets, in France, Italy, Germany and the US, among others, which could be exploited in such a way, building up a critical mass of material that will enable new connections to be made. The data-silo mentality could be gently undermined once scholars can see their own construct as remaining identifiable, while at the same time greatly enriched. The infrastructure will be sustained initially by King's College London and the UK National Grid Service (NGS), and subsequently as part of the European infrastructure being developed by the DARIAH project funded by the EU FP7 programme.

Paper 2: Data and Code for Ancient Geography: shared effort across projects and disciplines

Tom Elliott
New York University
tom.elliott@nyu.edu

Sean Gillies
New York University
sean.gillies@nyu.edu

Pleiades (http://pleiades.stoa.org) gives scholars, students and enthusiasts worldwide the ability to use, create and share historical geographic information about the Greek and Roman World. Pleiades is a joint project of three organizations: the Institute for the Study of the Ancient World (New York University), the Ancient World Mapping Center (University of North Carolina at Chapel Hill) and the Stoa Consortium for Electronic Publication in the Humanities (University of Kentucky). Our goal is a continuously updated, authoritative digital gazetteer for the ancient world, supporting the widest possible range of third-party digital projects and publications through open, standards-based interfaces. From its earliest concept days, Pleiades was intended to be broadly collaborative: employing, modifying and engendering open-content information and open-source software to accomplish its mission. This paper reports on the associated provisions and assesses their reach and effects within our user community, and beyond.

Pleiades employs a community-oriented, transparent editorial process that cultivates both contributions and critiques from the widest possible range of contributors. We aim to echo in a virtual environment the essential processes, workflows, resources and modes of interaction used by scholars to advance understanding of ancient sites, landscapes and geographic phenomena, but we also strive to open this environment to the widest possible range of interested participants. The bar to initial participation is purposely set low to encourage participation by scholars, students and enthusiasts alike, regardless of their degree status, institutional affiliation or skill level. All that is required is a verifiable email address and acceptance of a contributor agreement governing issues of professionalism, mutual respect, intellectual property, editorial policy, assertion of identity and the open-licensing of content.

Pleiades content combines “pure” data components (e.g., geospatial coordinates) with the products of analysis (e.g., toponymic variants with indicia of completeness, degree of reconstruction and level of scholarly confidence therein) and textual argument (e.g., comments and reviews). In part because of this hybrid constitution—and the varying definitions of intellectual property and “database rights” in differing legal jurisdictions—we have elected not to seek or assert any intellectual property ownership in the content on behalf of the project and its supporting institutions. Rather, our contributor agreement assumes (and contributors must affirm) that any IP rights inherent in the content remain with the contributors, who grant to the project (and therefore to its users) a Creative Commons, Attribution, Share-alike license, which permits reuse, redistribution and remixing of the content under clearly defined terms. These terms ensure the widest possible range of reuse (see below), without the need for copyright clearance requests and the like, while guarding against trivial reorganization and restrictive repackaging of the content that might inhibit such reuse.

On the software front, Pleiades is entirely open source. We make use of a number of externally developed components, and have contributed code to some of them. The Pleiades team has also created and released (under open-source licenses) a number of original components. Some of these are already in use beyond Pleiades, and we have received contributions of code from third-party developers for some of these.

- OpenLayers (http://openlayers.org) is the leading open source web map toolkit. Pleiades has modestly enhanced its features and employs it to provide contextual maps in its web application.

- Plone (http://plone.org) is a leading open source content management system. Pleiades has made modest improvements to its vocabulary manager and to its user interface framework, and contributed these code improvements back to the Plone code base.

- zgeo.* is a suite of Python software packages including: zgeo.geographer, zgeo.spatialindex, zgeo.atom, and zgeo.kml. These packages provide support for the Pleiades Entities component and enjoy contributions from programmers employed by The Open Planning Project and Makina Corpus SA.

Shapely and Rtree are general purpose Python GIS software that support the zgeo.* packages. Shapely enjoys contributions from programmers and researchers employed by Camptocamp SA, the University of California, and the National Oceanographic and Atmospheric Administration.

GEOS and SpatialIndex are low-level libraries for geometry and spatial indexing computing. Users include the PostGIS project and Autodesk. Pleiades has made modest contributions to each and helped SpatialIndex become an openly developed project.

Pleiades’ openness is driven in part by our sustainability plan and in part by the potential for re-use. Our initial content encompasses the compilation materials of the Classical Atlas Project, a 12-year, 200-person international collaboration that culminated in the publication of the *Barrington Atlas of the Greek and Roman World* (RJA Talbert, ed., Princeton, 2000). It is clear that no academic center or institution could maintain sufficient staff over the long-term to curate, maintain, update and diversify this unique assemblage of geospatial coordinates, toponymic records, temporal indicia and bibliographic citations drawn from a wide range of specialist literatures and primary sources. Consequently, we have embraced the necessity of pushing out the responsibility of (and opportunity for) creating and updating this content to interested parties across the discipline of ancient studies and beyond. Where these parties are employed, professional academics, we are gambling—it is true—on their willingness (and the willingness of their host institutions) to absorb the redistributory costs of content creation and maintenance as part and parcel of day-to-day research, publication and scholarly communication in their field. We are conscious that traditional “metrics” of reward (hiring, tenure, promotion) have not yet been adjusted to address such multi-institutional, asynchronous and piece-wise collaboration, and so we view our effort and our community as pioneers. Consequently, we endeavor to surface the details of individual contribution wherever possible: on profile pages, on individual records, and in the change histories that underly each such record.

The potential for reuse of Pleiades content is broad, and we hope to see both unexpected and serendipitous reuse cases arise from outside the project team. Our early collaborators are interested in a range of useful applications that can be foreseen or are already in prototype. One chief class of use is as an “authority list” for Greek and Roman geographic names and the locations associated with them. Both existing and new databases and digital resources can make use of Pleiades content (and our stable URLs for discrete elements therein) to refer unambiguously to the places and spaces mentioned in ancient texts, the subjects of modern scholarly works, the minting locations of coins, and the findspots of inscriptions, papyri, and the like. Using Pleiades as a central geographic authority reduces opportunities for ambiguity (consider that we know of 19 distinct cities named Apollonia in antiquity), while setting up the possibility of cross-project services and data sharing that exploits the common standard. In addition to the human-readable HTML interface, Pleiades provides access to its content in simple, standard formats that can be harvested, or aggregated dynamically, to produce dynamic maps using third party tools and services like Google Earth or Yahoo! Maps. We are also working with other projects to develop standards-based mechanisms for cross-project geographic search (e.g., relevant information within 30km of a named place). These same functional components will make it possible for scholars and students alike to pull Pleiades content into their own research and teaching tools and contexts, using them to solve problems, explore possibilities and produce map visualizations (mash-ups) for further sharing, reuse and publication.

Paper 3: Homer Multitext - Nine Year Update

C. W. Blackwell
Furman University
christopher.blackwell@furman.edu

D. N. Smith
Holycross
dnsmith.neel@gmail.com

In 2000, Gregory Nagy, in a review of Martin West's 1998 edition of the Homeric Iliad, contrasted the traditional critical editions of West and others with the ancient diorthosis (a word translated, perhaps loosely, as "edition") of the Alexandrian scholar Aristarchus: "I submit that Aristarchus' ancient edition of the Iliad, if it had survived in its original format, would in many ways surpass West's present edition. It would be a more useful—and more accurate—way to contemplate the Iliad in its full multiformity."

This assertion was the origin of the Homer Multitext, an effort to bring together a comprehensive record of the Homeric tradition in a digital library. This paper will describe the state of this project as it approaches the end of its first decade; this paper is not a progress report, but a description of an infrastructure that for both technological and what we might call semantic integration. In both its collection of data and its development of tools and infrastructure, the HMT has focused not on building a single-purpose application to support a particular theoretical approach, but on defining a long-term generic digital library expressly intended to encourage reuse of its contents, services, and tools.

Homeic scholarship in the 18th and 19th centuries was firmly based on the texts that survived through a manuscript tradition, most notably the great Byzantine codices of the Iliad and its ancient commentaries, the Homeric scholia. The 20th century saw the increasing recovery and publication of even older fragments on papyrus, but otherwise moved from a manuscript-based scholarship toward the scholarship of the critical edition. So, for example, the 1870s and 1880s saw the publication by W. Dindorf and E. Maas of editions of scholia organized by manuscript, the A and B manuscripts from Venice, and the T manuscript from London; in 1901 D. Comparetti edited a photo-facsimile edition of the A manuscript from Venice (part of an ambitious series of facsimile editions that was abandoned unfortunately incomplete due to lack of interest and funding). The 20th century saw the publication of a critical edition of the Iliad by T.W. Allen in the 1930s, and another by M. West in the 1990s, as well as a voluminous edition of the Homeric scholia by H. Erbse in the 1960s. In the cases of both the poetic text of the Iliad and the scholiasts' commentaries, these 20th Century publications are works of selection and aggregation, seeking to present a unified text of these ancient works, representing the best judgement of the editors.

Subsequently, however, scholarly assumptions in many circles about the nature of the poem and its commentaries have changed, and the range of questions that scholars would ask of these texts has expanded. The very existence of variation in the text has become a matter of historical interest (rather than a problem to be removed). The precise relationship between text and commentary, as expressed on the pages of individual manuscripts, hold promise to shed light on the tradition that preserved these texts, the nature of the texts in antiquity, and therefore their fundamental nature. We have found that the 20th century models of critical text-plus-apparatus is incapable of answering many of these new questions.

The best scholarly environment for addressing these questions would be a digital library of facsimiles and accompanying diplomatic editions. This library should also be supplemented by other texts of related interest such as non Homeric texts that include relevant comments and quotations and other collections of data and indices. Thus our focus on both collection of data and on building a scalable, technologically agnostic, infrastructure for publishing collections of data, images, texts, and extensions to these types. This infrastructure accomplishes retrieval and linking through abstract citation.

(This work is complemented by, and has been progressing in collaboration with, the work on Homer in the Papyri, which is also building a collection of diplomatic editions, to be supplemented by translations and commentaries, on papyrus fragments of epic poetry.)

We have presented aspects of this collection and infrastructure at previous meetings of the Digital Humanities Conference. This paper will summarize the project's goals, but focus on recent developments, specifically the ongoing publication of the Homeric Scholia, developments in our network services (specifically the third version of the Canonical Text Service and the RefIndex service, both of which now exist as Java Servlets and as Python applications running in the Google AppEngine), and our end-user application, a web-based interface to this library called "Pandect".

Neel Smith has been compiling editions of the Homeric scholia according to "new" principles that closely follow the real evidence for these ancient commentaries. Smith's edition acknowledges that each of the Byzantine codices in effect contains many discrete texts. The Venetus A, for example, contains a text of Proclus's Chrestomathy, the text of the Iliad, summaries of the books of the Iliad, at least four distinct scholiastic texts (as identified by their placement in discrete locations on each folio) and later notes and emendations. By describing these contents as separate texts, and by using a system of canonical citation to refer to portions of each text, and by using indices to associate these texts with the collection of folio-sides that constitutes this manuscript, we can approach the Venetus A as both a single artifact and as a notional "library" of texts. By virtue of our FRBR-like citation format, the CTS-URN, we can make general statements about passages from any of these texts, while also retraining the ability to treat each instantiation separately, as when a scholion appears in almost (but not quite) the same form on the A manuscript and the T manuscript.

The services that make this possible have been in development for years and are now ready for use in production. We have implementations of the Canonical Text Services protocol—for discovery and retrieval of text by means of arbitrary citations—in Java as a Servlet, to be run under Tomcat or Jetty, and as a Python application that can run in Google's AppEngine space. Likewise the RefIndex service, which permits generic access to indices that allow simple pairings between texts (at any level from the text-group, or author, level down to the citation level or a specified substring), objects in a collection (such as a collection of morphological data, a lexicon, or a collection of manuscript folios), or images.

These two implementations allow us to offer these services through local servers, for the greatest flexibility, or through Google's service, for global access and greatest reliability.

Finally, Pandect is our open-source web-based application for accessing the materials of the HMT. Its main function is to mediate between the user and the network services of the HMT, and as such is should be an entirely generic tool, useful for any other project that implements the CTS protocol.

By virtue of this citation+service approach to our digital library, Pandect can discover texts, data, and images; it can also provide basic navigation and manipulation. It is not, however, merely a multi-column viewer for text and images. Each instance of the application is based on a Scenario, which defines relationships between collections of texts, collections, and images. The scenario might know that a collection contains data about manuscript folios, and that these in turn are related to images and to xml texts. The user's experience, then, consists of navigating a digital library in which each object in view knows its relationship to all others. Navigation of one object will percolate across all others. At any given point, the user's current view—for example, a Homeric texts, a scholion, and images of two folios—is preserved, can be addressed, and can be exported as a simple XML expression of a directed graph (using the GraphML schema). Because each of these objects is identified through canonical citation, these digraphs capture the relationships among scholarly objects; they can serve not only as bookmarks to the state of a browser, but as independent objects of analysis, aggregation, or manipulation.

After nine years of development, we hope to make the case that the Homer Multitext has not only produced a large body of valuable data, but also a robust body of source code that could be broadly useful to the community of digital humanists. Its approach to primary sources—favoring diplomatic editions and facsimiles wherever possible—intends to invite the widest possible scope for re-use of its data. Its emphasis on simple indexing rather than complex and specialized internal markup is based on the assumption that it requires less knowledge to integrate texts with simple markup and simple, documented indices, than to disaggregate an elaborately marked up texts that embeds links to other digital objects. The HMT's emphasis on canonical citation insures that its contents can continue to interrelate with each other, can be abstracted, and be re-used into the future. And our emphasis on services defined by documented protocols should allow the HMT to advance in functionality and reliability, and should allow other projects to draw on the HMT's contents through a consistent interface that is independent of any specific technological implementation.

For a Dynamic Model of Textual Variation: What Do We Need?

Chair: Dino Buzzetti
University of Bologna

The idea of organizing a joint discussion on a dynamic model of textual variation stems from the observation that previous work carried out independently by Malte Rehbein and the Signum centre in Pisa could be related in a more comprehensive framework. Malte Rehbein had shown that the chronological succession of the variations introduced in the handwritten records of the statutes of the city of Göttingen in the course of the 15th century could be established by means of structured external information stored in a database. In their turn, results obtained at the Signum centre on the semantic structure of Giordano Bruno's *De gli eroici furori*, showing the diachronic variation across successive parts of the work, could be analysed through the textualization of a comprehensive structural representation. So it was possible to observe that the process of textual variation would refer to semantically structured information on the one hand, and that the transition across distinct interpretational variants would invoke a textualized representation of its overall process on the other. And from these mutual observations a tentative model of textual dynamics could be elicited.

By a joint presentation of three related papers we intend to provide no more than a simple illustration of the problems to be solved in order to build a viable model of textual dynamics and variation, without presuming that the tentative solutions here proposed would be the only possible ones, not even to say the best. We would only like to contend that finding an effective solution to these problems would render digital editions a far more appealing means of representing and studying textual materials than their conventional counterparts.

A further observation is here in order. In their case studies, Malte Rehbein and the Signum team shall in turn present a model of textual and interpretational variation, but each applied to a different text. In the final paper I shall try to show how these respective models can be related in an overall model of textual variation and I shall try to provide an example of its application to the same text by referring to a very simple textual fragment.

Paper 1: Topic Maps and MVD for the

representations of interpretative variants

Alida Isolani
isolani@signum.sns.it
Scuola Normale Superiore

Claudia Lorito
Chiara Genovesi
Daniele Marotta
Marco Matteoli
Cinzia Tozzini
Signum, Centro di ricerche informatiche per le discipline umanistiche

1. Introduction and aims

Signum, in collaboration with the Istituto Nazionale di Studi sul Rinascimento, gained a remarkable experience in analysing and processing humanistic texts, which resulted in the development of a search engine for XML documents: TauRo-core.

In this research field a demand has emerged for a system aimed at developing semantic research and aids for the reading and interpretation of texts.

The fundamental assumptions of our work have been proven consistent with the research project carried out by Prof. Dino Buzzetti about the connection between expression and meaning in texts [BUZ, 2002]. This perspective prompted us to reach a further objective, i.e. testing — through an algorithmic approach — the links which exist between different interpretations of the same text, in order to create an output format which is analogous to the rendering of textual variants.

2. The application

Our research focused on the analysis of *De gli eroici furori* — a Renaissance Italian text written by Giordano Bruno in 1585 — and achieved two results:

- the creation of a system for facilitating semantic research and text reading [BUZ, 2004];
- the creation of a graph of interpretative variants in MVD format.

2.1. The system for facilitating reading and semantic search

We accomplished an online system that allows a guided reading and a semantic search of De gli eroici furori.

This text has been chosen, firstly, because of its philosophical nature, which allows the outlining of complex and articulated conceptual structures. Moreover, the subdivision of the text into various dialogues enables the reader to follow the diachronic development of concepts. Therefore, *De gli eroici furori* is the ideal subject for our case study (as previously shown by Bassi et al. project [BDEL, 2007]).

In order to realise our prototype, we decided to delineate three interpretative variants concerning a single conceptual unity (**intellect**) in three dialogues: I, 1; I, 4; II, 5 (this selection is aimed at showing the development of this conceptual unity across the text). A representative example concerns the relation between love and intellect: love enlightens intellect (I, 1); intellect is metaphorically defined as the faculty of sight (I, 4), so it leads to the vision of the highest good (II, 5), the true object of love (I, 4).

> I, 1 - p. 792: ... l'amore non è cieco in sé, e per sé non rende ciechi alcuni amanti... l'amore illustra, chiarisce, apre l'intelletto...
>
> I, 4 - p. 819: ... l'amor eroico per quanto tende al proprio oggetto ch'è il sommo bene; e l'eroico intelletto che gionger si studia al proprio oggetto che è il primo vero o la verità absoluta...
>
> I, 4 - p. 827: ...perché o significa la potenza visiva, cioè la vista, che è l'intelletto...
>
> II, 5 - p. 956: ...nove ciechi... sentiro aspergere dell'acqui bramate, aprîro gli occhi e veddero...[la luce] che sola possea mostrargli l'imagine del sommo bene...

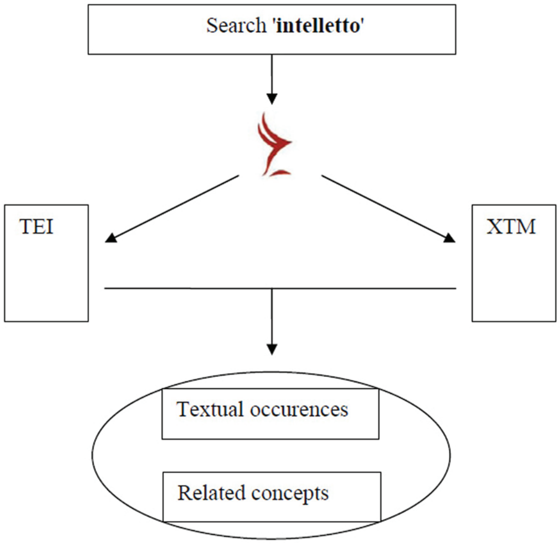


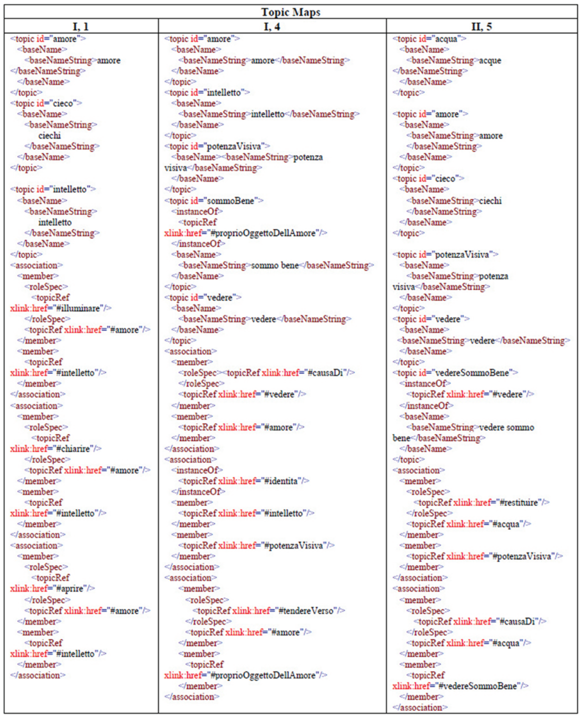
Figure 1. System architecture

To represent the semantic dimension of *De gli eroici furori* topic maps have been chosen, whose features are similar to those pertaining to the natural process of text interpretation; in particular, the standard language adopted to write topic maps is XTM. This system allows a guided reading by means of a schematic visualisation of the topic maps relating to the conceptual unity in question. Semantic search is carried out by the TauRo-core search engine. When a query is submitted, TauRo-core searches in *Eroici furori* text encoded in TEI format and searches in topic map files encoded in XTM format. The two results are then conveyed into a single visualisation, in order to make available to the user textual occurrences as well as "conceptual constellations" that are related to the query (Fig. 1).

2.2. The graph of interpretative variants

Many Digital Humanities IT applications focused on textual variants, leaving aside interpretative variants. Our aim is to contribute to overcome this discrepancy by the adoption of a unified format consistent with both types of variants, i.e. MVD.

This attempt is made possible by the fact that different interpretations are described through a XTM standard, which — being a XML language — is itself a text. Consequently, Desmond-Schmidt's algorithm can be applied to represent any textual interpretation as a path spanning from the first to the last node in an oriented MVD graph.



Topic Maps		
I, 1	I, 4	II, 5
<topic id="amore"> <baseName> <baseNameString>amore </baseNameString> </baseName> </topic> <topic id="cieco"> <baseName> <baseNameString> ciechi </baseNameString> </baseName> </topic> <topic id="intelletto"> <baseName> <baseNameString> intelletto </baseNameString> </baseName> </topic> <association> <member> <roleSpec> <topicRef xlink:href="#illuminare"/> </roleSpec> <topicRef xlink:href="#amore"/> </member> <member> <topicRef xlink:href="#intelletto"/> </member> </association> <association> <member> <roleSpec> <topicRef xlink:href="#chiarire"/> </roleSpec> <topicRef xlink:href="#amore"/> </member> <member> <topicRef xlink:href="#intelletto"/> </member> </association> <association> <member> <roleSpec> <topicRef xlink:href="#aprire"/> </roleSpec> <topicRef xlink:href="#amore"/> </member> <member> <topicRef xlink:href="#intelletto"/> </member> </association>	<topic id="amore"> <baseName> <baseNameString>amore</baseNameString> </baseName> </topic> <topic id="intelletto"> <baseName> <baseNameString>intelletto</baseNameString> </baseName> </topic> <topic id="potenzaVisiva"> <baseName><baseNameString>potenza visiva</baseNameString> </baseName> </topic> <topic id="sommoBene"> <instanceOf> <topicRef xlink:href="#proprioOggettoDellAmore"/> </instanceOf> <baseName> <baseNameString>sommo bene</baseNameString> </baseName> </topic> <topic id="vedere"> <baseName> <baseNameString>vedere</baseNameString> </baseName> </topic> <association> <member> <roleSpec><topicRef xlink:href="#causaDi"/> </roleSpec> <topicRef xlink:href="#vedere"/> </member> <member> <topicRef xlink:href="#amore"/> </member> </association> <association> <instanceOf> <topicRef xlink:href="#identita"/> </instanceOf> <member> <topicRef xlink:href="#intelletto"/> </member> <member> <topicRef xlink:href="#potenzaVisiva"/> </member> </association> <association> <member> <roleSpec> <topicRef xlink:href="#tendereVerso"/> </roleSpec> <topicRef xlink:href="#amore"/> </member> <member> <topicRef xlink:href="#proprioOggettoDellAmore"/> </member> </association>	<topic id="acqua"> <baseName> <baseNameString>acque </baseNameString> </baseName> </topic> <topic id="amore"> <baseName> <baseNameString>amore </baseNameString> </baseName> </topic> <topic id="cieco"> <baseName> <baseNameString>ciechi </baseNameString> </baseName> </topic> <topic id="potenzaVisiva"> <baseName> <baseNameString>potenza visiva</baseNameString> </baseName> </topic> <topic id="vedere"> <baseName> <baseNameString>vedere</baseNameString> </baseName> </topic> <topic id="vedereSommoBene"> <instanceOf> <topicRef xlink:href="#vedere"/> </instanceOf> <baseName> <baseNameString>vedere sommo bene</baseNameString> </baseName> </topic> <association> <member> <roleSpec> <topicRef xlink:href="#restituire"/> </roleSpec> <topicRef xlink:href="#acqua"/> </member> <member> <topicRef xlink:href="#potenzaVisiva"/> </member> </association> <association> <member> <roleSpec> <topicRef xlink:href="#causaDi"/> </roleSpec> <topicRef xlink:href="#acqua"/> </member> <member> <topicRef xlink:href="#vedereSommoBene"/> </member> </association>

Table 1. The topic maps of the above-mentioned textual passages from the three dialogues.

Let us see how this idea has been applied to our case study by means of a concise example focused on the three textual passages mentioned above and drawn according to the topic maps of the three dialogues (table n. 1).

From these three topic maps results a MVD graph, a simplified (as far as nodes are concerned) version of which is reproduced below, in order to show some macro-variants:

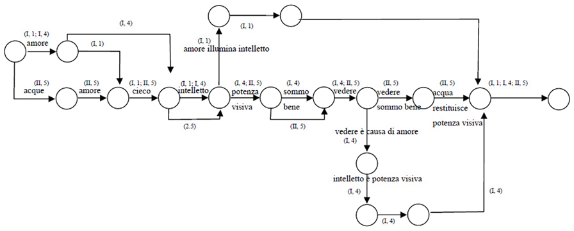


Figure 2. MVD graph

3. Conclusions

This project is in line with the theoretical view according to which it is possible to pass, in the digital processing of humanistic texts, from the expressive to the interpretative level. It was carried out through the creation of a tool for semantic search and guided reading, and through the implementation of an algorithm to create a data structure that we can visit with an analogous algorithm used for textual variants (i. e. MVD).

In the future, our purpose is to use the representation of interpretative variants by MVD graph to redefine the expressive level of texts, in compliance with the scheme suggested by Prof. Buzzetti.

4. Bibliography

[BDEL, 2007] S. Bassi, F. Dell'Orletta, D. Esposito, A. Lenci. *Computational linguistics meets philosophy: a Latent Semantic Analysis of Giordano Bruno's texts*, RINASCIMENTO XLVI, 2007, pp. 631 - 651.

[BUZ, 2002] D. Buzzetti. *Digital Representation and the Text Model*. New Literary History 33(1), 2002, pp. 61 - 88.

[BUZ, 2004] D. Buzzetti, *Diacritical Ambiguity and Markup*, in D. Buzzetti, G. Pancaldi, and H. Short (eds.), Augmenting Comprehension: Digital Tools and the History of Ideas, London-Oxford, Office for Humanities Communication, 2004, pp. 175-188

[BM, 2006] D. Buzzetti, J. McGann. *Critical Editing in a Digital Horizon*, in *Electronic Textual Editing*, ed. L.

Burnard, K. O'Brien O'Keeffe, and J. Unsworth, New York, The Modern Language Association of America, 2006, pp. 51 - 71.

[BUZ 2009] D. Buzzetti, *Digital Editions and Text Processing*, in M. Deegan and K. Sutherland (eds.), *Text Editing, Print, and the Digital World*, Aldershot, Ashgate, 2009, p. 45-62.

[MESC, 2005] F. Meschini, *Le mappe topiche: come imparai a non preoccuparmi e ad amare i metadati*, Associazione italiana biblioteche. Bollettino AIB, 2005, n. 1 p. 59 - 72.

Paper 2: Multi-Level Variation

Malte Rehbein
National University of Ireland, Galway
malte.rehbein@nuigalway.ie

The paper is a follow-up to a joint presentation by Dino Buzzetti and myself at the Digital Humanities conference in Oulu 2008[1]. There, we applied a case study of a complex medieval text to generate a general model towards dynamic editions. It lies in the nature of such a conference that a 20 minutes presentation is hardly enough to achieve both: an introduction to a project whose textual variation is vital in understanding its content and genesis, and to derive a general view of both its markup and its structural information. One of the reviewers commented at that time, already before the conference: "I also feel far too much is being proposed here for one twenty-minute talk; either part I or part II would still be straining the limits of what can fit". This paper, proposed for the Digital Humanities conference 2009 will thus concentrate on the case study *kundige bok*, but bring in new results from recent research on the text and its variants.

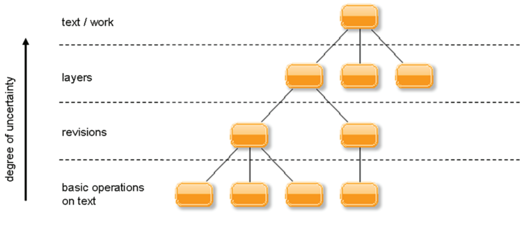


Figure 1: Operation-Revision-Layer-Model of textual genesis (draft).

Kundige bok is a 15th century legal text manuscript, characterised by many revisions where the different layers of the text usually represent a new stage of the town law and must thus be treated equally[4].

The digital edition of *kundige bok* tries to achieve a dynamic approach, allowing the user to generate views of the text in the status it had at a certain point in time. Variation in the case of kundige bok is three-fold:

1. The textual variation, determined by the operations performed on the text by the medieval scribes.

2. The sequence of revisions / layers of the text as a whole respectively the underlying law reflecting its genesis. This can be quite linear but also very complex as for instance Peter Robinson has shown in his work on Chaucer's Canterbury Tales.

3. The editor's or user's interpretation of this sequence. Owing to the fact that we often cannot reveal the actual development of the text due to our limited knowledge about its production, this leads to an interpretational variation.

While we can be quite sure about the textual variation itself, the more we try to group these operations on the text to revisions and text layers, the more our understanding of the text becomes inconclusive and uncertain (see figure 1). This is illustrated by an example which is also used in the following to explain the three kinds of variation that can be found in *kundige bok* and can be represented using graph structures, similar to the works by e.g. Huitfeldt and Sperberg-McQueen [2] or Schmidt[3].

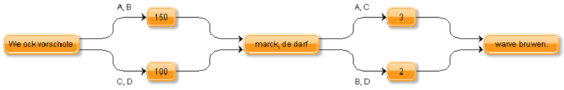


Figure 2: Text variants as graph structure.

The example, taken from *kundige bok*, is a statute about beer brewery. What we find nowadays looking at the manuscript, is the sentence "We ock vorschote 100 marck, de darf 3 warve bruwen" with the numbers '100' and '3' struck out and the numbers '150' and '2' written above the line. This allows four different variants of the regulations as shown in figure 2 using a graph representation. Here, I use the vertices as a container for the portions of the text and the edges to describe the variants. Each variant (A-D) is thus represented as a path through the graph.

However, we do not know per se which of these variants really have existed as town law and in which order. In other words: the grouping of the operations on the text (e.g. the replacement of '100' by '150) to revisions and the (chronological) ordering is ambiguous. As long as we have no further information, we must live with three different possibilities through the genesis of the text, each of those representing a different grouping of the operations on the text to revisions and different layering. Comparable to the example of textual variants, this kind of interpretational variation can be represented by a graph structure as well (figure 3). The three paths (a)-(c) through the graph represent the possible evolutions of the text. Note, that there is no edge between B and C since this would be possible only if one of the changes was reversed.

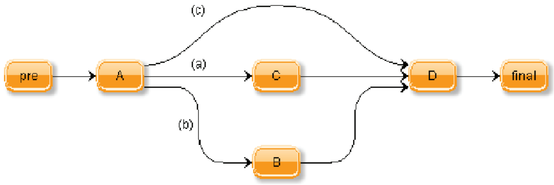


Figure 3: Interpretational variation of layering.

This paper discusses furthermore the question of whether a graph representation like this and the underlying mathematical model is indeed suitable to deal with the uncertainty that forced its creation. Although we do not have evidence that one of the three paths actually took place, we might have a clue which makes one or two more probable than the others. The extension of the directed graph in to a weighted directed graph could be a way to deal with this additional information. The weights, put on the edges can indicate which variation is more likely than the other and easily be changed should new information be available (figure 4).

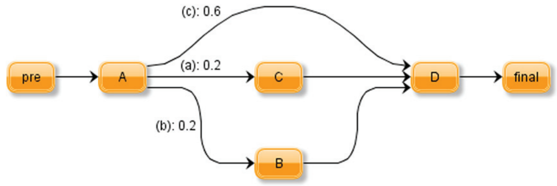


Figure 4: Interpretational variation as weighted graph.

This data can then be used for various purposes: firstly, to create a dynamic visualisation of the different variants, e.g. using colours for the probabilities, secondly to automatically process the data, e.g. ignoring paths with low probabilities. Finally it might be a good approach, especially when you have a lot of such cases for which you can give probabilities individually but not for the evolution of the text as a whole. Applying weights to all single variations can then lead to an overall view by mathematical calculation, e.g. computing the "cheapest" path through the graph, an algorithm related to the Travelling Salesman (TSP) problem.

Other examples in *kundige bok* are, of course, more complex and this was a major issue in analysing the text and creating the edition and its user-interface. The question of uncertainty arises generally in the creation of genetic editions (and these thoughts are also inspired by the working group on genetic editions within the TEI Special Interest Group on Manuscripts[6]), thus, dealing with the described issues might be a step forward towards a model for encoding and formalisation of genetic editions as well

References

[1] D. Buzzetti and M. Rehbein, *Towards a Model for Dynamic Editions*, paper given at the Digital Humanities conference, Oulu 2008.

[2] C. M. Sperberg-McQueen and Claus Huitfeldt, *GODDAG: A Data Structure for Overlapping Hierarchies, in: DDEP-PODDP 2000*, ed. P. King and E.V. Munson, Lecture Notes in Computer Science 2023 (Berlin: Springer, 2004), pp. 139-160, online: http://www.w3.org/People/cmsmcq/2000/poddp2000.html.

[3] http://multiversiondocs.blogspot.com.

[4] M. Rehbein, *Reconstructing the Textual Evolution of a Medieval Manuscript*, in: Literary & Linguistic Computing, forthcoming.

[5] http://en.wikipedia.org/wiki/Traveling_salesman_problem.

[6] M. Rehbein and J. Tonra, *Encoding Genetic Editions — Two Case Studies*, paper given at the TEI Members Meeting, London 2008.

Paper 3: How to Build a Textual Data Model ?

Dino Buzzetti
University of Bologna
buzzetti@philo.unibo.it

The purpose of considering a type of representation of textual variation based on external structured information, on the one hand, and a type of representation of external interpretational variants based on an a textualization of their comprehensive description, on the other, aims at finding proper ways of building an overall *data model* for the processing of textual information of both kinds, the representational (its expression) and the interpretational (its content) one. For we think that sheer representation, without processing, leaves a *digital* (i.e. processable) edition essentially incomplete.

The basic principle of interconnection between the "image" or expression of the text [Segre, 1985: 378] and its meaning or information content can be stated as follows: "there may be different ways of understanding what is said and different ways of saying what is meant," or, to put it in other words, "the fixity and invariance of the expression (or the content) respectively entail the indetermination and variance of the content (or the expression)." [Buzzetti, 2004: 180] As referred to structural representations both of the the textual data (in terms of "syntactic markup structures") on the one side, and of the corresponding information content (in terms of "objects, properties, and relations" of specific semantic domains) on the other, [Dubin, 2003: 2] the same principle can be expressed by affirming "that the same markup can convey different meanings in different contexts," and "that markup can communicate the same meaning in different ways using very different syntax." [Dubin and Birnbaum, 2004: 1] Are there means of connecting the two sides in a systematic way?

In his *Kundige bok* edition Malte Rehbein has shown that it is possible to produce a two-tiered representation, a marked-up transcription of the text and its variants, and a database, that connects the definition of the several text layers with further contextual information. Both representations have an operational character, so that changes in the the markup modify the database entries, whereas the database entries produce newly marked-up views of the text. In their turn, the Signum group at the Scuola Normale Superiore in Pisa have shown that a topic map representation of the content of Bruno's *Eroici furori* can be connected with a joint visualization of textual occurrences and their semantically related notions. A change in the topic map produces a reorganization of the intratextual relations, and in reverse a new organization of textual relations produces a different topic map.

Moreover, both case studies have implemented a comprehensive representation of the variants, respectively textual and interpretational, by means of the Multi-Version Document data structure (MVD) introduced by Desmond Schmidt. [Schmidt and Colomb, forthcoming] The MVD data structure is a directed graph with a start and an end node and each textual layer or interpretational description is represented by a different path on the graph. A topic map representation can be serialized through the XTM standard and treated exactly in the same way as a textual document. Both the editorial and the interpretive practice imply a one-to-many relation between the different textual or interpretational versions and their comprehensive representation, or "logical sum" [Thaller, 1993: 64 and Buzzetti, 2002: 77-79], although in reverse order, as the editor goes to one single reconstruction from the many witnesses of the text, and the literary critic from one edition to its many interpretations. And the MVD data structure seems to provide a reliable and processable representation of the *logical sum* of both textual and interpretational variants.

The two case studies deal with different texts, but by using the same kind of representation for both textual and interpretational variants of the same text, we think of simplifying the task of mutally mapping them onto each other according to the concept of a dynamic model that can be elicited from the ambivalence of markup [cf. Buzzetti, 2004 and Buzzetti, 2009]. As a diacritical mark, markup is ambiguous and can be seen both as the value and as the rule of a structuring operation. [Buzzetti and McGann, 2006: 67-68] By applying this principle to both the textual and the interpretational representations, we can be able to use either of them as a set of instructions to restructure and reorder the other one. A polysemic textual fragment can be easily used to demonstrate the working of such a model.

Can the MVD data structure provide a sound and effective basis for mapping corresponding alternative paths from the textual variation graph to the interpretational variation graph and vice versa? The question is open. But finding a proper data model for a viable computational solution to this kind of mapping raises a crucial challenge to any attempt of building a comprehensive data model for textual information. And we hope that providing a closer illustration of the problem may assist in finding a solution.

Biliography

[Buzzetti 2002] D. Buzzetti, *Digital Representation and the Text Model*, in « New Literary History » 33:1 (2002), pp. 61-87.

[Buzzetti 2004] D. Buzzetti, *Diacritical Ambiguity and Markup*, in D. Buzzetti, G. Pancaldi, and H. Short (eds.), *Augmenting Comprehension: Digital Tools and the History of Ideas*, London-Oxford, Office for Humanities Communication, 2004, pp. 175-188.

[Buzzetti, 2009] D. Buzzetti, *Digital Editions and Text Processing*, in M. Deegan and K. Sutherland (eds.), *Text Editing, Print, and the Digital World*, Aldershot, Ashgate, 2009, pp. 45-62.

[Buzzetti and McGann, 2006] D. Buzzetti and J. McGann, *Critical Editing in a Digital Horizon*, in L. Burnard, K. O'Brien O'Keeffe, and J. Unsworth (eds.), *Electronic Textual Editing*, New York, The Modern Language Association of America, 2006, pp. 51-71.

[Dubin 2003] D. Dubin, 'Object mapping for markup semantics,;' in B.T. Usdin (ed.), *Proceedings of the Extreme Markup Languages 2003 Conference* (Montreal, Quebec, August 2003) <http://www.idealliance.org/papers/extreme/proceedings/xslfo-pdf/2003/ Dubin01/EML2003Dubin01.pdf> (26 October 2008).

[Dubin and Birnbaum: 2004] D. Dubin and D. Birnbaum, 'Interpretation beyond markup,' in B.T. Usdin (ed.), *Proceedings of the Extreme Markup Languages 2004 Conference* (Montreal, Quebec, August 2004), <http://www.idealliance.org/papers/extreme/proceedings/xslfo-pdf/2004/Dubin01/EML2004Dubin01.pdf> (26 October 2008).

[Schmidt and Colomb, forthcoming] D. Schmidt and R. Colomb, 'A Data Structure for Representing Multi-version Texts Online,' *International Journal of Human-Computer Studies*, forthcoming.

[Segre, 1985] C. Segre, *Avviamento all'analisi del testo letterario*, Torino, Einaudi, 1985, p. 378.

[Thaller, 1993] M. Thaller, "Historical Information Science: Is There Such a Thing? New Comments on an Old Idea," in T. Orlandi (ed.), *Discipline umanistiche e informatica: Il problema dell'integrazione*, Roma, Accademia Nazionale dei Lincei, 1993, pp. 51-86.

Digital Editions, Past and Future

Chair: John Lavagnino
National University of Ireland & King's College London

Introduction

Our panel surveys the field of digital editions, ranging from past achievements and future projects. Our focus is on the problem of shared infrastructure: the lack of reusable machinery for producing and publishing such editions is the biggest obstacle to their creation and survival. As long as each project is obliged to develop much of its own infrastructure, we won't see speedy progress.

But building such infrastructure is not an easy task. Scholarly editions are complex publications: they need to present documents and works with a high degree of fidelity to original sources, and add on top of that an analytic layer that describes variations and offers discussions of difficult words or passages. Tools that are flexible enough for this are hard to build; this session describes some, but they do not exist in large numbers. And the complexity and the limitations of software for digital repositories and digital libraries suggest that in this more complex arena it's a challenge to devise general solutions.

Our demands are also increasing. We are now dissatisfied by editions that are docu-islands; we want everything to link together, better than print editions can do it. We want an appealing and consistent reader experience, not something that each edition does differently. We want continuous development of a broad collection of editions, one that many scholars can contribute to, and not a set of closed projects that don't talk to one another.

The first talk, by John Lavagnino, is about the past: it describes the problems of digital scholarly editing in the light of a completed project, with particular attention to the problem of handling the large number of types of text an edition might include, and the problem of building machinery that has to keep working if the edition is to remain available.

The second talk, by Fotis Jannidis, is about the immediate future: it describes the tension between the complex requirements of a digital genetic edition of Goethe's *Faust* and the difficulties of fulfilling them with a working environment which has to be developed or at least heavily adapted for this task.

Finally, Susan Schreibman's talk looks at the current state of international efforts to develop infrastructure further, and at initiatives to build a digital text-editing community in Ireland based on those efforts.

Paper 1: Experiences of the previous generation: the Thomas Middleton edition

John Lavagnino
National University of Ireland & King's College London
John.Lavagnino@kcl.ac.uk

I am one of the general editors of a collected edition of Thomas Middleton's works, published in 2007—more than ten years behind schedule. Of course, one of the most common questions we heard during that period was: why was it taking so long? (The other common question was: why don't you add more features and more material?) This talk attempts to explain.

Of course, it may just have been too ambitious a schedule to begin with; most comparable editions have taken at least as long as ours did. Other people have found the same task difficult: at least two substantial projects to produce a collected Middleton were started during the twentieth century and never finished; the last published collection appeared in 1885–6. And, as with most large-scale projects, ours would have gone better with more funding.

But we thought we had reasons to believe it would go faster than it did. From the start, we did not intend to pursue historical and bibliographical research as far as possible: collecting the research of the last few decades and making it available would have been enough, and while there is much new work in the edition we did not try to push as far as possible with every single play. We collected a large group of editors so that few contributors were working on more than one or two works. And, with some exceptions that I'll mention, this edition did not intend to be particularly innovative in its digital methodology. It's true that it was always going to produce both print and digital editions, not especially common in the early 1990s; but when we began our idea of a digital edition was of something you could feed to a concordance program, not a full-scale online publication.

I think we did largely succeed in addressing delays deriving from the *traditional* research. Any group of more than seventy collaborators will have some whose lives obstruct their work in some way, but most were able to complete their portion on schedule and our other delays made it possible to get everything else completed. We were doing less than is common for comparable editions today, as a comparison with Fotis Jannidis's talk in this panel will show: we only prepared and published an edited text, and didn't include diplomatic transcriptions of the witnesses as well, something commonly felt to be important for digital scholarly editions even in the early 1990s.

We also did offer slightly more digital support for the work of preparing the edition than is normal for print-oriented editions: I devised a system to allow editors to prepare their notes using the line numbers of a draft printing of the text; it then automatically adjusted them to the right line numbers for the final printing, in which textual corrections usually made all the line numbers somewhat different. The normal approach has been for a staff of student employees to make these adjustments manually, but we didn't have such a staff, and in a large edition it is a large task (our final text has over 60,000 references by line number). The system was a good deal of work to build, but it had unexpected side effects, since a system to adjust line numbers in this way is also a system to check line numbers and lemmata for correctness. It was perhaps the only really innovative element of our digital approach.

Because our edition offered only the edited text, we did not run into the problem of relating it to the sources: so we largely avoided one of the usual encoding problems. This is the difficulty arising from the fact that a scholarly edition has to do more than represent the works included in an adequate way, possibly covering numerous genres. It also needs to represent variation in texts in those genres, something that is challenging with XML, and indeed is often managed using non-XML systems, such as version control software. Ideally, the system should be able to relate the variation to the document structure, but this is difficult with non-XML approaches.

But if the Middleton edition avoided some difficult problems, it still faced the software problem. With pre-digital scholarly editions, one of the biggest costs came from typesetting; that problem is transferred to software and its application in the digital era, whether the output is print or digital. But this isn't just a shifting of costs. In a pattern visible in many areas of the digital humanities, there is a shift: from a process that ends with a static product, supported by established traditions for distribution and preservation, to the creation of a machine for generating what readers see. The digital edition is more like a customized car than a book, and keeping it running is a serious problem, as it requires continued attention and vigilance in obtaining replacement parts.

Though the Middleton edition is entirely in TEI XML, it has actually been published only on paper (so far), and so we might think has avoided the software problem at the expense of missing out on online publication. But to preserve the possibility of updated editions and selective reprints, the machinery that made the printed pages must be preserved, and it is quite as complicated as a system for online publication. Perhaps more, because print makes some demands that online publication does not raise: fitting as much as possible onto each page, for example, in our case by printing the text in two columns and the notes in three. Scholarly editions are recognized as posing typesetting problems that are often quite complicated; that doesn't get easier because software is involved.

Once we've built this machinery for a single edition, could it not then be reused for many? We would get more of a return from our effort, and there would be more interested parties to support the infrastructure. In part this has been done: much of the underlying typesetting work, to print a text with line numbers and notes keyed to those line numbers, is done by a layer added on top of Donald Knuth's TeX typesetting system that Dominik Wujastyk and I wrote, EDMAC, and this component has been used not only by the two of us but by numerous other editions. We don't get support from those users, but of course just being able to list them is a credential we can draw on in seeking funding, and the possibility of a shared infrastructure is there.

There are two major problems to be solved in this area, though. One is the update problem: all software systems either change or die, and the infrastructure we created has not been changing. TeX is one of the most stable systems there is, but even in the TeX world there have been changes, mostly in font handling, that we haven't been keeping up with.

The bigger problem is that of the variety of texts. Even within the Middleton edition there are many works that required specialized handling. I don't just mean ones where we chose an editorial method different from that used for the rest of the works: the occasional cases where we print multiple versions, or the one case of a parallel text. I mean cases where the work itself required unusual typesetting because that's what the work was: one short theological work that used a six-column parallel layout, for example, or one place where very long marginal notes are suddenly necessary (in a passage not even by Middleton: it's by Ben Jonson, one of five contributors to an elaborate pageant). A general framework that can handle the whole variety observed in texts needs to be very general indeed; and asking the author to rewrite the work to make it more tractable is not an option.

We went part of the way towards this goal in EDMAC: my field is English and Dominik's is Sanskrit, and we discovered that we shared almost no basic assumptions about what an edition is, with the useful result that we built something very generic. It is only dealing with a few aspects of scholarly editions, though: the numbering of lines and the creation of multiple series of notes keyed to line numbers. Even in this restricted domain there are practices we did not reproduce: such as one found in the Monumenta Germaniae Historica series, of numbering not only the lines of edited text on each page, but also the lines in the footnotes. (Page images available at the Monumenta Germaniae Historica digital web site illustrate the phenomenon.)

In the world of digital curation, the software problem is familiar, even if more attention has been given to more fundamental problems about static data formats. (Hunter, 2006, is one work that does address software issues, and the MLA's guidelines for scholarly editions also suggest attention to the problem.) All the more reason, then, why a shared infrastructure is desirable, rather than one that is unique to each edition; but the difficulties of creating a generally-useful system are large, and much more work is needed to overcome them.

References

Hunter, Jane (2006). Scientific Publication Packages—A Selective Approach to the Communication and Archival of Scientific Output, 1. First published 18 November 2006: http://www.ijdc.net/ijdc/article/view/8.

Knuth, Donald (1984). *The TeXbook*. Reading, Massachusetts: Addison-Wesley.

Lavagnino, John, and Wujastyk, Dominik (1996). *Critical Edition Typesetting: The EDMAC format for Plain TeX*. San Francisco: TeX Users Group.

Modern Language Association, *Guidelines for Editors of Scholarly Editions*, version of 25 September 2007. http://www.mla.org/resources/documents/rep_scholarly/cse_guidelines (accessed 14 November 2008).

Monumenta Germaniae Historica digital. http://www.mgh.de/dmgh/ (accessed 14 November 2008).

Taylor, Gary, and Lavagnino, John, general editors (2007). *Thomas Middleton: The Collected Works* and *Thomas Middleton and Early Modern Textual Culture: A Companion to The Collected Works*. Oxford: Clarendon Press.

Paper 2: Requirements and tools for a modern genetic edition of Goethe's *Faust*

Fotis Jannidis
Technische Universität Darmstadt
sprachli@linglit.tu-darmstadt.de

The Goethe–Schiller archive in Weimar, the Goethe-Haus in Frankfurt and my group in Darmstadt are preparing a digital critical edition of Goethe's drama Faust; work started at the beginning of 2009. Astonishingly enough, there is no modern critical edition. Thus scholars still have to use the edition created 100 years ago as part of the *Weimarer Ausgabe* of Goethe's work, which makes an understanding of the existing material and the genetic processes rather difficult. Goethe worked for fifty years on the drama, and about 1000 manuscript pages exist, documenting this long genesis; most of these pages are part of *Faust II*, which Goethe finished a short time before his death.

The new edition of *Faust* will allow access to all manuscripts as facsimile and as diplomatic transcription. In this kind of edition the facsimile and transcription should be presented side by side; moreover, in contrast to the manner of a number of such editions currently on the web, the transcription should be closely linked to the facsimile, first by preserving many aspects of the look of the facsimile and by being linked bidirectionally to it on a micro level. It should also be possible to view the facsimile alone and render transcriptions of verses as mouse-over activated tooltip; or, vice versa, to view solely the transcription and additionally view parts of the facsimile.

A detailed genetic analysis of each document and of the text will allow viewing the genetic process of a document and of structural units like verse or scene. The archival (or document-centric) view of the material is thus complemented by a conceptual view of the text as a drama and its units.

These basic requirements determine the architecture of the markup, which will be mostly layered stand-off markup to avoid problems with overlapping hierarchies and to allow easier processing. (On the problems with the markup of genetic editions, see: Vanhoutte, 2006; Van Hulle, 2006.) To name some of these layers:

- **Linking facsimile and text**
  This can be done using the facsimile markup in the TEI P5 guidelines.

- **Diplomatic transcription**
  Features of the manuscripts that must be recorded include the direction of the text, shift of hand, textual alterations, gaps, non-textual marks on the page etc. Some of these features are covered by the TEI guidelines; some of them, which describe image-like qualities of a manuscript and of the writing, have to be expressed by other means such as SVG.

- **Genetic markup**
  This is an interpretation of the document which is described by the facsimile and the transcription under a genetic perspective. At the moment no markup schema for this type of edition exists, but there is one under development by a working group which is part of the TEI Special Interest Group on Manuscripts. The genetic markup will be split into different levels: a functional analysis of textual alterations, sequences and time lines, grouping changes—like revisions on a page or changes of conceptual units of the text, a coordinate system which links together chunks of text which belong together, a way to add the interpretation which is the basis for a markup decision, and uncertainty.

One of the problems for an edition like this is the lack of editors and other tools which are easy for scholars to use and allow them to concentrate on the editorial work; and which are robust and integrate easily with other tools. The project TextGrid, which has reached early beta status after three years of development, has as its aim the creation of such an environment for text-centered studies. If its second work phase is funded, it will offer some basic tools, but all tools have to be adapted for the Faust edition to allow editors to enter the different layers of markup, and new tools will have to be integrated into the workflow. The special requirements of each digital edition seem to make it impossible (or at least unaffordable) to offer a complete solution which allows their creation, change and enhancement, and publication out of the box. Therefore some time of an editorial project has to be spent on setting up a workflow and overcoming the problems of data conversion between different tools or merging the outcome of different tools. As the ease of use of tools has a direct impact on the speed with which a digital edition is created, a solid knowledge of existing tools and of the difficulties to adapt them is crucial for a new project, especially since up to now there have not been enough reviews of specialized tools in the digital humanities. (And there is almost no culture of publication for the concepts and algorithms of specialized tools for the digital humanities, either.)

The talk will concentrate on the requirements for the

Faust edition and the problems of finding and adapting tools and making them interoperable; as such it offers an interesting parable on the beauty of theory and the pitfalls of reality.

References

TEI Special Interest Group on Manuscripts. http://www.tei-c.org/wiki/index.php/SIG:MSS (accessed 14 November 2008).

TextGrid. http://www.textgrid.de/ (accessed 14 November 2008).

Vanhoutte, Edward (2006). Prose Fiction and Modern Manuscripts: Limitations and Possibilities of Text Encoding for Electronic Editions. In Burnard, Lou, O'Keefe, Katherine O'Brien, and Unsworth, John (eds), *Electronic Textual Editing*. New York: Modern Language Association, pp. 161–180.

Van Hulle, Dirk (2006). Authorial Translation: Samuel Beckett's *Stirrings Still / Soubresauts*. In Burnard, Lou, O'Keefe, Katherine O'Brien, and Unsworth, John (eds), *Electronic Textual Editing*. New York: Modern Language Association, pp. 150–160.

Paper 3: An E-Framework for Scholarly Editions

Susan Schreibman
Digital Humanities Observatory
susan.schreibman@gmail.com

As discussed in our introduction, one impediment to more scholars creating digital scholarly editions is the size of the technological hurdles that need to be overcome. The Digital Humanities Observatory (DHO), a newly-founded national digital humanities centre located in Dublin, Ireland, but serving the higher education sector for the island of Ireland (both North and South), has begun a process of consultation with its academic partners and other interested parties (publishers, university administrators, and scholarly societies) to develop a framework for digital scholarly editions.

It almost goes without saying, or at least it should, that not everyone who wishes to edit an edition for electronic publication necessarily wants to become a digital humanist. Despite the fact that there is a large and established community of practice around editing texts utilizing TEI/XML, there does not yet exist a framework for digital scholarly editions. The majority of scholarly editions are still developed as one-off productions with the content tightly integrated with the software. We also don't, as a scholarly editing community, have agreed-upon formats, protocols, and methodologies for digital scholarly editing and editions. Moreover, many of the more mature first-generation digital projects creating online editions from print sources have more in common with digital library projects—i.e. editions created with a light editorial hand, minimally encoded and with little more contextualization than their print counterparts.

The DHO is entering into this endeavour on the back of several European projects that have begun to explore creating a framework for digital scholarly editions. Two of them are especially important for the framing of the DHO's project: TextGrid and Interedition:

- TextGrid is a German-based project funded by the Bundesministerium für Bildung und Forschung that has developed a research infrastructure for collaborative textual processing in the text sciences and the humanities in general. TextGrid, like the DHO, is based on grid technology.
- Interedition is a newly-formed COST Action funded by the European Union. It is a project initiated by the Huygens Institute (The Netherlands) to enhance the international digital infrastructure for scholarly editorial work. In a series of meetings, this project brings together researchers in the field of literary research and IT to develop a roadmap for creating a shared supranational networked infrastructure for digital scholarly editing and analysis.

Building on the experiences learned from these projects, the DHO has begun a series of dialogues to help establish a set of protocols, methodologies, rights management and technical procedures to create a shared infrastructure for digital scholarly editions in Ireland.

This paper will report on two of the major activities held in Spring 2009: a one-day symposium bringing together academics, publishers, librarians, and technologists; and a Spring Scholarly Editing School. The Spring School (to be held in April), will bring together many noted scholarly editors and theorists of editing (both digital and print). The week-long school will build on the preliminary work of the symposium to help establish, on the one hand, system requirements that the technical staff of the DHO can implement, and on the other an economic, social, and cultural framework so that scholars can receive the training to work in this environment, the academic credit for their scholarly work, and the peace of mind in knowing their work will be part of an established infrastructure.

References

Anonymous (2008). *Supporting Digital Scholarly Editions: A Report on the Conference of January 14, 2008*. http://www.virginiafoundation.org/NEH%20Workshop%20Report%20FINAL-3.pdf (accessed 14 November 2008).

Cummings, James (2007). The Text Encoding Initiative and the Study of Literature. In Siemens, Ray, and Schreibman, Susan (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell, pp. 451–76.

Digital Humanities Observatory. http://dho.ie (accessed 14 November 2008).

Finneran, Richard J. (ed) (1996). The Literary Text in the Digital Age. Ann Arbor: University of Michigan Press.

Interedition. http://www.interedition.eu/index.php/Main_Page (accessed 14 November 2008).

McGann, Jerome (2002). Literary Scholarship in the Digital Future. *Chronicle of Higher Education*, 13 December: B7–B9.

Price, Ken (2007). Digital Scholarly Editions. In Siemens, Ray, and Schreibman, Susan (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell, pp. 434–50.

Schreibman, Susan (2002). Computer-mediated Texts and Textuality: Theory and Practice. *Computers and the Humanities*, 36: 283–93.

TextGrid. http://www.textgrid.de/ (accessed 14 November 2008).

Critical Code and Software Studies

Chair: Marc Marino
University of Southern California

For the future of Digital Humanities, software cannot remain either a black box or some sacred text that only the high priests can access. Software intersects with and coordinates so many aspects of our lives, and a few lines of source code can determine processes all around us. In 2008, The University of California San Diego organized the *Softwhere Studies* workshop to further the development of Software Studies, the critical analysis of the cultural circulation of computer software. Critical Code Studies is a subset of Software Studies that uses critical theory to explicate the extra-functional significance of computer code, exploring not merely what the code does but what it means. This panel offers a series of talks that take up a Software Studies or Critical Code Studies approaches to demonstrate the role of reading code and software using the interpretive approaches of the humanities. Such moves will lead not only to the tearing down of disciplinary boundaries but also a vital understanding of the way software and source code structure our political and personal lives. After a capsule definition of critical code and software studies, panelists will present their papers.

Paper 1: Expressive Processing

Noah Wardrip-Fruin
University of California, Santa Cruz
nwf@ucsd.edu

The forthcoming book *Expressive Processing* (MIT Press, 2009) uses its title's term to point toward two important issues for humanities scholars.

First, "expressive processing" encompasses the fact that the internal processes of digital media are designed artifacts, like buildings, transportation systems, or music players. As with other designed mechanisms, processes can be seen in terms of their efficiency, their aesthetics, their points of failure, or their (lack of) suitability for particular purposes. Their design can be typical, or unusual, for their era and context. The parts and their arrangement may express kinship with, and points of divergence from, design movements and schools of thought. They can be progressively redesigned, repurposed, or used as the foundation for new systems - by their original designers or others - all while retaining traces and characteristics from prior uses.

Second, unlike many other designed mechanisms, the processes of digital media operate on, and in terms of, humanly-meaningful elements and structures. For example, a natural language processing system (for understanding or generating human language) expresses a miniature philosophy of language in its universe of interpretation or expression. When such a system is incorporated into a work of digital media - such as an interactive fiction - its structures and operations are invoked whenever the work is experienced. This invocation selects, as it were, a particular constellation from among the system's universe of possibilities. In a natural language generation system, this might be a particular sentence to be shown to the audience in the system output. From the output sentence it is not possible to see where the individual elements (e.g., words, phrases, sentence templates, or statistical language structures) once resided in the larger system. It is not possible to see how the movements of the model universe resulted in this constellation becoming possible - and becoming more apparent than other possible ones.

Paper 2: What Counts as Code to Criticize? Interpreting flow control and natural language programming

Jeremy Douglass
University of California, Santa Barbara
jeremydouglass@gmail.com

Sites of interpretation in the digital humanities range from the signifiers at the software interface all the way down to the electromagnetic phenomenology of hardware mechanisms. Critical Code Studies situates interpretation at the site of reading and writing source code. These human-machine interlanguage exists both between and prior to the hardware and its interface, often with many other interpretive sites above (application, OS, emulator, etc.) and below (assembly, machine code, etc.). As interlanguage, code carry a variety of philosophical presumptions, aesthetic overdeterminations, and historical traces that may benefit from exegesis. But what counts as code to criticize? This presentation considers critical interpretation in relation to two cases of code that differ substantially from the paradigmatic use of languages such as C/C++ and the attendant assumptions about the relationships between code, programmer, compiler, and software. The first case, natural language programming, emphasizes the expressive power of code and its accessibility to non-programmers, while it simultaneously risks obscuring the deep structures that connect compiled code to software praxis. The second case, flow control programming, uses visual spatial relationships rather than text as the top-level paradigm for specifying software action. Both cases expand our idea of what aspects of rhetoric and visual culture might be brough to bear in humanistic code criticism.

Paper 3: Hacktivism and the Humanities: Programming Protest in the Era of the Digital University

Elizabeth Losh
University of California, Irvine
lizlosh@uci.edu

Despite the dismay of university administrators, who are often hesitant to recognize computer programming as a form of campus free speech, students continue to deploy code to further activist agendas   Yet media reports about hacktivist activities by undergraduates and graduate students rarely place these computer-coded expressions of protest in the context of other kinds of campus activism or critical engagement. Whether the student is generating electronic boarding passes to protest Homeland Security policies or data mining for self-interested Wikipedia edits by corporations and politicians, the attention goes to the programmer's identity as a hacker rather than as a student. Although Siva Vaidhyanathan and others have issued a manifesto to promulgate "critical information studies" as an interdisciplinary field of study that could serve as the logical successor to the areas of academic inquiry that arose from the protests of the nineteen-sixties and seventies, which is "needed to make sense of important phenomena such as copyright policy, electronic voting, encryption, the state of libraries, the preservation of ancient cultural traditions, and markets for cultural production," advocacy for these issues in the university setting does not achieve the kind of visibility that was associated with previous movements that assembled crowds of individuals for face-to-face interactions in physical public space to achieve an end to the Vietnam War, milestones on civil rights issues, affirmative action, or divestment in South Africa. This presentation builds upon the models of critical information, critical code, and software studies to examine the analytical frameworks that might reinvigorate the program of the political explication of coding spaces.

Paper 4: Explorations of a Terrorist Expert System (2 parts)

Mark Marino and Stephanie August

Critical Code Studies: The Terrorist Database

Mark Marino
University of Southern California
markcmarino@gmail.com

In the midst of the cold war, the U.S. Department of Defense was beginning its efforts to track terrorists using databases enhanced with scruffy AI. Early prototypes attempted to model how terrorists shared knowledge in order to determine how these organizations might take advantage of associations. While scholars have analyzed the textual trail of terrorists and counter-terrorism efforts, examining the code of one of these early models reveals the ways in which terrorists were conceived in this time period. This paper uses the methodology of Critical Code Studies to show what this LISP source code and design documents say about the cultural moment in which the military attempted to model geopolitical insurgencies.

How Cold War Computers Assign Blame: An exploration of why machines think more but not more deeply.

Stephanie August
Loyola Marymount University
saugust@lmu.edu

Data is inherently dynamic. If the failures of the past few decades (or human history) prove nothing else, they underscore the foolishness of only tracking and making decisions based on only one aspect of data. Today's databases still suffer from a rigidity that handicaps their ability to perform as "expert" systems. Early prototype databases modeled terrorist social networks and made inferences about their associations using relatively straightforward rules and representations. Inheritance relations of object-oriented programming have found their way into relational databases, yet our representation mechanisms and query languages have note changed substantially in twenty years. While search and retrieval mechanisms have become more sophisticated, software programs still do not reflect understanding of the documents they process. And adapting them to changing values and attitudes is problematic. The future promises the opportunity to express preferences over hard and fast rules and the ability to reprogram aspects of our code without recoding and recompiling an entire software system. Until then, our ability to reason about our data, while more extensive, won't be much different in terms of depth and nuance than it was in the 1980s and will still be bound by the static nature of the knowledge we store in our programs. Decisions made on the advice of these expert systems (from military to medical realms) will most certainly suffer.

Preserving Virtual Worlds: Models & Community

Chair: Jerome McDonough
University of Illinois, Urbana Champaign

The Preserving Virtual Worlds project, a research initiative funded by the Library of Congress' National Digital Information Infrastructure for Preservation Program, is currently investigating issues surrounding the archiving of computers games, virtual worlds and interactive fiction, and developing technical solutions and recommendations on practices to support the long-term preservation of these unique cultural works. Participants include the Rochester Institute of Technology, Stanford University, the University of Illinois at Urbana-Champaign, and the University of Maryland.

This session will present results from some of the initial research conducted as part of this project, with a focus on re-examining existing conceptions regarding the operations of digital archives with respect to both:

- the relationships between archivists and the communities they serve; and

- the applicability of traditional (and new) models of bibliographic and archival description with respect to archiving software generally, and games in particular.

The papers will focus on three key subjects: (1) the ontological issues surrounding the definition of a computer game and the problems they present for description using models such as the entity-relationship model set forth in the Functional Requirements for Bibliographic Records Final Report; (2) the Open Archival Information System's notion of a designated community and examine whether it is an appropriate model for approaching the archiving of games, given both the costs associated with such archiving and the pre¬¨existence of preservation activities within the gaming community; and (3) the problems presented by the existence of an extremely heterogeneous community for dealing with issues surrounding documentation of both software and hardware components of games and their use context.

All authors have indicated their willingness to participate in this session.

Paper 1: Twisty Little Passages Not So Much Alike: Applying the FRBR Model to a Classic Computer Game

Matthew Kirschenbaum
University of Maryland
mgk@umd.edu

Doug Reside
University of Maryland
dreside@umd.edu

Neil Fraistat
University of Maryland
fraistat@mac.com

Jerome McDonough
University of Illinois, Urbana Champaign
jmcdonou@illinois.edu

Dennis Jerz
Seton Hill University
jerz@setonhill.edu

Humanities scholars have continually confronted questions regarding the boundaries of the texts that they study and the complex inter-relationships that can exist among various editions, translations, and printings —in short, the versions—of a work. While librarians have long recognized the distinction between a work as an intellectual creation and its embodiment within a particular physical form (and the need to adequately describe both), the publication of the Functional Requirements for Bibliographic Records Final Report by the IFLA Study Group on the Functional Requirements for Bibliographic Records (FRBR) marked a pronounced increase in the level of attention that the library community has devoted to these issues. In the decade since the Final Report was issued, a tremendous amount of discussion has occurred regarding FRBR's interpretation and its appropriate application within bibliographic systems. At the same time, there has been almost no cross-communication between humanities scholars engaged in the kind of work described above ("textual studies" as it is called) and library specialists.

As extraordinarily complicated as the relationships between the various versions of a particular text can be within the world of traditional manuscripts and print publications, the move to electronic text, and in particular highly interactive texts such as interactive fiction and computer games, has rendered these relationships even more vexed and difficult to describe adequately. Because each individual or subsequent encounter with the same interactive work can generate different output texts, the adequacy of traditional descriptive models applied by librarians to enable scholars' access to textual materials needs to be carefully examined.

A fundamental component of any effort to preserve digital resources is the development of systems to describe and track the components of a digital work and to relate works (and their physical embodiments) to each other, including describing the provenance of manifestations as a work evolves over time. As a test of existing library practices, our project has been examining the application of the FRBR entity-relationship model to computer games and interactive fiction, including the seminal work ADVENTURE. This paper will examine the difficulties encountered by the project in seeking to apply the FRBR entity relationship model within the realm of computer games, and our project's suggestions for "pretty good" practices for the application of FRBR and traditional bibliographic descriptive practices to this ever-evolving electronic genre.

ADVENTURE, aka ADVENT or Colossal Cave, was first written and programmed by Will Crowther in 1975, and then revised and extended by Don Woods tn 1977. It offered an underground fantasy world inspired by Crowther's experience as a caver in Kentucky,Äôs Mammoth Cave. ADVENTURE is a not a virtual world merely in a casual, metaphoric sense, but in explicit formal terms: the program contains a parser and a world-model that tracks the state of the electronic space, for example what objects are and are not in a user-avatar's possession and where the avatar is in a global environment (Montfort 2003). While ADVENTURE was originally played on a teletype machine, it achieved widespread popularity and recognition when it was released onto the nascent ARPANET. Its historical and cultural significance is captured in Richard Powers's novel Plowing the Dark (2000) where the author describes a generation of computer users united by their shared experience of the Colossal Cave's iconic virtual landmarks. This alone would make ADVENTURE an extremely strong candidate for the preservation work we are funded to undertake. However, the recent recovery of the original version of the game from backup tapes at Stanford University—an event which received widespread notice around the Web—lends the work fresh interest and appeal (Jerz 2007).

After Woods published his significant revision to Crowther's original source code, numerous additional development forks complicated ADVENTURE's composition history. The game has also been ported and migrated across nearly every platform and operating system that is extant. ADVENTURE can be played on an Apple II and it can be played on the latest Mac OS. It can be played on Linux systems and Windows. It can be played on an iPod. Early on Microsoft released its own proprietary version of the game, adding an extra room to the underground caverns. Different scoring systems have evolved, and fans can wrangle endlessly over which is to be considered canonical. ADVENTURE thus presents itself as a rich and complex digital object, not only for its internal workings as a functioning program but as a cultural artifact that has been continually reimplemented and reinterpreted by a mature fan community.

FRBR is, at first glance, a promising mechanism for representing this heritage. It is an entity-relationship model capable of discriminating among changes to the substance or "content" of the work, as well as it's physical embodiment in particular carrier media. In a traditional FRBR representation, one might start with the **work** that is *Hamlet*. The different versions of the play that are extant are the work's **expressions**. These expressions are realized in **manifestations**, i.e. the folios and quartos that have survived, as well as the more modern editions based upon those sources. A discrete artifact that one holds in hand, for example the copy of the Arden Shakespeare sitting nearby on my bookshelf, is an **item**. At the same time, certain long-standing challenges present even with more traditional applications of the FRBR model. For example, there is no formal consensus on how much of the work has to change before a new expression is declared. Catalogers (for FRBR is primarily a cataloger's tool) are asked to rely upon common sense, community practice, and other heuristics.

In the case of an electronic object, the complications proliferate almost exponentially (Renear 2006). At first it might seem that all versions of ADVENTURE should be the "Work," a particular instance of the game (the last version modified by Don Woods, for instance) should be the "Expression," a particular file with a unique MD5 hash should be the "Manifestation," and an individual copy of that file (perhaps on a Commodore 64 664 Block disk) would be the "Item." But what if the text read by the reader is exactly the same, but the underlying code is different? These variants might be simple (a non-compiled comment added to the Fortran code), peripheral (such as the ability to recognize "x" as a synonym for the command "examine"), or very large (a port of the code from Fortran to BASIC). Should these code level variants be considered different expressions? To further complicate matters, what if the Fortran code was exactly the same but compiled to two different chips? For example, an IBM mainframe and a Commodore 64 might both have a Fortran compiler, but the two compilers will interpret the Fortran to a different set of set instructions. It might also be the case that two Fortran compilers designed by different compilers will generate slightly different machine language. Should these compiled executables, different in their binary structure but based on the same Fortran, represent different "Manifestations" or different "Expressions"?

Finally, even two files with exactly the same MD5 signature participate in a larger software environment at runtime. The drivers that run the video interface, the keyboard, the memory, and the disk drives arguably become part of ADVENTURE when the user is playing the game. For instance, the experience of playing the game using the 6507 chip in a Commodore 64 hooked up to a black and white television may be different than the experience of playing the game on the same chip in a Commodore SX64 (the all-in-one machine some felt fit to call "portable"). Should the software environment on which the binary is executed be a part of the classification scheme at all?

We have applied the FRBR model to three different and specific instances of ADVENTURE: the source and data files as retrieved on April 27, 2008 at 6:01 pm from Dennis Jerz's server (http://jerz.setonhill.edu/if/crowther/), the DOS Windows executable of these files edited to compile under GNU g77, a free FORTRAN compiler (http://www.russotto.net/%7Erussotto/ADVENT/), and a pirated copy of "Apple Adventure" on a 5 1/4" diskette in one of the project members' personal possession. This work will be presented in the course of the paper, together with rationale and discussion in the context of the kind of issues enumerated above. We will also discuss the significance of this work for the broader digital humanities community, especially in so far as it represents the intersection of library and information science, textual studies, and software forensics.

As more and more libraries and repositories begin the process of collecting born-digital object, they will invariably encounter material that transcends the boundaries of documents, email, and other more or less conventional forms of electronic records. ADVENTURE, as both a working computer program and as a virtual world, as well as an artifact with widespread popular interest, is a harbinger of the kind of content which increasingly need to be accessioned, cataloged, and described. FRBR represents the library community's best effort to date to distinguish between different versions and editions of a work. We believe the work discussed represents an important test case for FRBR's applicability to complex born-digital objects.

References

Functional Requirements for Bibliographic Records Final Report, IFLA (1998): http://www.ifla.org/VII/s13/frbr/frbr.htm.

Jerz, Dennis (2007). "Somewhere Nearby is Colossal Cave: Examining Will Crowther's Original 'Adventure' in Code and in Kentucky." *Digital Humanities Quarterly* 1.2: http://www.digitalhumanities.org/dhq/vol/001/2/000009.html

Montfort, Nick (2003). *Twisty Little Passages: An Approach to Interactive Fiction*. Cambridge: MIT Press.

Powers, Richard (2000). *Plowing the Dark*. Farrar, Strauss, Giroux.

Renear, Allen. "Is An XML document a FRBR Manifestation or a FRBR Expression? ‚Äî Both, Because FRBR Entities are not Types, but Roles." *Proceedings of Extreme Markup Languages 2006*: http://idealliance.org/papers/extreme/Proceedings/html/2006/Renear01/EML2 006Renear01.html#tod0e5

Paper 2: The Open Archival Information System Reference Model vs. the BFG 9000: Issues of Context and Representation in Game Software Preservation

Henry Lowood
Stanford University
lowood@stanford.edu

Jerome McDonough
University of Illinois, Urbana Champaign
jmcdonou@illinois.edu

In February of 1993, the Committee for Film Preservation and Access submitted a statement to the National Film Preservation Board of the Library of Congress advocating the creation of a national policy to preserve the nation's motion picture heritage, and arguing that such a policy would "be incomplete—utterly pointless—unless there is a guarantee of access to the films that are being preserved," (CFPPA, 1993). This argument was summed up in the title of their statement, a phrase that has echoed throughout the library preservation world during discussions of efforts to preserve digital information: "Preservation Without Access is Pointless."

While insuring access to preserved material is a necessary condition of preservation, we believe that efforts to preserve computer games demonstrate definitively that it is not a sufficient condition for their scholarly use. Scholars require more than simply access to cultural materials; they require resources for understanding them as fully as possible, which is a far more difficult goal to achieve. Games are very complicated technological artifacts, and as Lowood (2008) notes, they are also incredibly complicated cultural artifacts. Understanding a game requires extensive documentation of their technical nature, of their use, and just as important, of the contexts for their use.

Issues surrounding the extent and nature of documentation necessary to preserve digital information have been extensively studied and debated by a variety of communities within the past decade. One of the significant results of these discussions has been the emergence of *The Reference Model for an Open Archival Information System* (OAIS) (CCSDS, 2002) as a foundational standard for the design and operation of digital archives. OAIS provides both a functional model for the operation of an archive as well as an information model describing the necessary types of information that must be acquired and maintained to preserve digital objects and their relationships to one another. While originally crafted to inform the efforts of space data archives, OAIS has gained acceptance far beyond its original community and is being used to assist in the design of preservation repositories by the international digital library community.

The information model set forth by the OAIS reference model is predicated on the idea that preserving any form of digital content will also require the preservation of a body of additional knowledge necessary to understand and interpret that digital content. This additional knowledge takes a variety of forms, but two are of particular significance to the preservation of computer games:

> *Representation Information*—formally defined as “information that maps a data object into more meaningful concepts,” representation information is that information which allows a series of zeros and ones constituting a digital data stream to be interpreted as meaningful. Representation information is of two types: structure information, which maps a bit stream into common data types such as character data, integers, arrays, etc., and semantic information, which allows a user to meaningfully interpret the data. Structure information tells you that a series of zeros and ones represents a four-digit integer; semantic information informs you that the four-digit number represents the date of publication for a work.
>
> *Context Information*—defined as “information that documents the relationships of the content information to its environment,” context information includes information documenting the circumstances in which information being archived was originally produced. It also documents relationships that may exist between a particular piece of content and other content within the archive or elsewhere. The OAIS reference model identifies provenance information as a specific type of context information, but it does not provide guidance about the nature or extent of the types of relationships between items in the archive (or outside it). This information is needed to preserve digital content that will carry meaning for future historians and other scholars.

The assumption underlying the requirement to store representation and context information is that preservation of digital content must entail preservation of the ability to fully interpret and understand content in the digital archive, both at a basic technical level (being able to render the digital data in a manner a human being can apprehend) and at a more sophisticated intellectual level (being able to understand the digital data,Äôs significance and relevance).

Archivists face a key question in trying to conform to the OAIS reference model. What is the nature and extent of the representation and context information they must maintain in order to allow users to fully understand the information being stored? Taken to its logical extreme, the semantic representation information for a digital version of Borges’ *El Jardín de Senderos que se Bifurcan* should include both a Spanish dictionary and grammar (and quite possibly bilingual editions of each to insure that it is possible to decipher Spanish without actually knowing the language to begin with). For complex data formats and data sets, the costs of acquiring and preserving a complete set of representation information could be extremely high. The OAIS reference model recognizes this dilemma, and states that the extent of the representation information maintained by an archive should be based upon the knowledge possessed by the designated community that the archive serves. If that community is fluent in Spanish, there is no need to maintain a Spanish dictionary and grammar in representation information set for a Spanish language work.

Now we come to the rub. With respect to context information, the OAIS reference model provides less guidance. Its emphasis on the origin of digital information, as well as on the relationship to other material in the digital archive, suggests that the concept of context information as developed by OAIS’s authors is at least somewhat informed by notions of archival bond and *respect des fonds* from archival theory (Gilliland-Swetland, 2000). By maintaining relationships between archival records that were originally established by their creators and documented by their methods of organizing these records, the archivist traditionally preserves the original context for each individual record. At its core, the archival profession assists users in understanding the totality of information in the archive. This approach works for many electronic records, or even scientific data sets, but it may not be appropriate for all digital media and all forms of digital information.

The Preserving Virtual Worlds project has been investigating how computer games and interactive fiction might be preserved in a manner consistent with the OAIS reference model. The basic notions of representation information and context information would seem to be applicable to the preservation of computer games as digital content. However, our research to date suggests that OAIS's assumptions regarding the relationship between archivists and the community they serve do not necessarily hold for a digital library of computer games (and probably for other complex forms of interactive media or software). The OAIS reference model seems to assume a relative homogeneity in the user community's knowledge base; when this is the case, archival training is geared toward determining the representation information necessary for members of that community. With respect to this model, our investigations suggest that potential users of a game archive demonstrate a challenging range of technical knowledge with respect to gaming technology and use. We have not found a common intellectual grounding in software design and implementation or in game design that would allow an archivist to readily discern an appropriate level of representation information to record for games.

The same pattern holds for context information. Allowing users of a game archive to fully understand the context for a particular game also requires providing them with much more information than the implicit information provided by digital records,Äô archival bond. But potential users of a game archive come from a variety of perspectives and with vastly differing research needs that require different contextualizing information. The needs of a game researcher investigating the relationships between game companies and user communities with respect to issues of game mods and intellectual property law are very different from the needs of a researcher investigating the influence of the development of pixel shading technology on game art, and both require significant information beyond copies of the games themselves to support their endeavors. These differing users bring varying levels of knowledge of game history and game play to their work which complicate the task of any archivist attempting to determine the forms and extent of context information that must be preserved along with the game itself.

In this paper, we will explore some of the records that will need to be preserved along with game software to address the diverse and important research questions that future historians will pose. If assumptions regarding the designated community for an archive break down, the application of OAIS,Äôs notions of representation information, context information and provenance information becomes highly problematic for our efforts to preserve computer games. Using id Software,Äôs historically important game DOOM as an example, this paper will discuss specific problems that have emerged in applying the OAIS information model to the archiving of computer games. We will also discuss ways digital game archivists might be able to circumvent these problems.

References

Borges, Jorge Luis (1942). *El Jardín de Senderos que se Bifurcan*. Buenos Aires: SUR.

Consultative Committee for Space Data Systems (CCSDS) (2002). *Reference Model for an Open Archival Information System* (OAIS). CCSDS 650.0-B-1 Blue Book. Washington, DC: CCSDS Secretariat: http://public.ccsds.org/publications/archive/650x0b1.pdf

Gilliland-Swetland, Anne J. (Feb. 2000). *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment*. Washington, DC: Council on Library and Information Resources: http://www.clir.org/pubs/reports/pub89/pub89.pdf

Lowood, Henry. "Found Technology: Players as Innovators in the Making of Machinima." *Digital Youth, Innovation, and the Unexpected*. Edited by Tara McPherson. The John D. and Catherine T.MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press, 2008. 165-196. doi: 10.1162/dmal.9780262633598.165: http://www.mitpressjournals.org/doi/pdf/10.1162/dmal.9780262633598.165

Paper 3: Game Change: The Role of Professional and Amateur Cultures in Preserving Virtual Worlds

Kari Kraus
University of Maryland
karimkraus@gmail.com

Rachel Donahue
University of Maryland
donahrm@gmail.com

Megan Winget
University of Texas
megan@ischool.utexas.edu

In 2006, Margaret Hedstrom and a team of researchers at the University of Michigan published the results of an experiment exploring how users evaluate the authenticity and usability of computer games and other digital materials that have been preserved through multiple methods (Hedstrom 2006). The study found that users tended to prefer playing emulated and migrated versions of a popular 1980s-era computer game known as Chuckie Egg to the original version. Arguing the importance of the user's perspective in archival decision-making, the authors begin and end the paper with a call for further research into the "needs and preferences" of the user community. The purpose of our presentation is to develop this line of inquiry by drawing on our experiences with the Preserving Virtual Worlds and Game Preservation projects.[1] Where we depart from Hedstrom and her colleagues is in how we position the user within the preservation system. While their assumption is that users are only or primarily consumers of digital information, ours is that they increasingly play an active role in its preservation. To frame the issue in archival terms, we see users taking responsibility for collecting, managing, and creating long-term access to computer games. Because our interest lies with the contact zone between hobbyists and professional archivists, we will enumerate and analyze how we are currently collaborating with or relying on the user community to preserve virtual worlds, with an eye to how these relationships might eventually be codified within a larger preservation framework.

Web 2.0 provides one obvious context in which to consider the rise of the amateur archivist. In recent years much has been made of the allegedly stark contrast between amateur and professional culture, and those writing about it tend to structure the conversation in markedly polemical terms. While proponents of amateur culture extol the democratization of the means of production, opponents denounce the decline of subject expertise, certified credentials, and standards. This paper attempts to shift the focus and substance of the debate in two ways: first, by framing amateur and professional culture in terms of content *preservation* rather than (or in addition to) the more customary content *creation* (user-*preserved* media rather than user-*generated* media). We will draw our examples from the creative/expressive domains of computer games, interactive fiction, and 3D virtual worlds. Second, by demonstrating the interconnectedness of professional and amateur efforts to preserve virtual worlds. We propose that digital preservation of virtual worlds requires cooperation among several different stakeholders, including users, designers, creators, and archivists.

A librarian with a book she's unsure of might turn to WorldCat for bibliographic information. Where do we turn for video games? There is no professional resource that can offer us the same depth and breadth of information for video games that WorldCat offers for more standard information resources. Instead, we must turn to the wealth of web sites maintained by the gaming community (MobyGames, Home of the Underdogs, and GameFAQs, to name a few). In the case of the PVW project, we are indebted to this community for providing us with fresh perspectives on our preservation practices, methodologies, and theories. On a purely practical level, we rely on the resources the community produces and manages to help us document context information, as well as prepare descriptive, technical, and administrative metadata for the digital objects in our case set. Moby Games, for example, is an online game database billed as "a historical archive, documentation, and review project for all electronic games (computer, console, and arcade)" with two overarching goals in mind: "1.) to store a diverse array of information about games, who created them, what their system requirements are, what their game screens look like . . . all browseable via an easy-to-use, hyperlinked interface. 2.) to let the public contribute to each entry in the database, whether it be with new entries, additional information, simple ratings, or detailed reviews." While the screenshots, descriptions, and game metadata available on the site are all useful, perhaps most valuable is the information on the hardware and software platforms necessary to play different versions of the games we are archiving. One such game is Robert Pinsky's *Mindwheel*, one of five interactive electronic novels published by Broderbund software as part of its text adventure series. Initially released for IBM and Apple, versions of the game were also adapted for Atari and Commodore systems.

The same caveats that apply to the reliability and accuracy of user-generated content in other research scenarios also apply here, and consequently we take pains to cross-reference many different sources of information. It should be noted, however, that quite often the most authoritative resources by conventional standards are authored by individuals who are themselves dependent on or contribute to the user-generated wikis, databases, and discussion boards highlighted here. For example, Wikipedia has been a regular port of call for learning about--in some cases identifying--the publishers and rights-holders of various games in our archiving case set. Pinsky's *Mindwheel* is instructive in this regard. *Mindwheel* was originally owned and developed by Synapse Industries (or Synapse Software Corporation) and published by Broderbund Software, who acquired the IP rights in c. 1984. The complex corporate history of Broderbund, whose creative assets have been repeatedly bought and sold over the past 20 plus years, is well documented on Wikipedia. The genealogy is as follows: The Learning Company purchased Broderbund in 1998; Mattel then purchased the Learning Company in 1999 and sold it to Gores Technology Group in 2000, who in turn sold the Learning Company's holdings to Ubisoft and Riverdeep in 2001. Mindwheel's corporate chain of transmission is thus at least six levels deep. Based on the Wikipedia entry, it appears that the Broderbund assets and IP rights are now under the control of Riverdeep. Attempting to verify this, we contacted the Riverdeep company, who informed us that the program was no longer available or supported, but that we should consult Moby Games for more information about the developer. Likewise, Nick Montfort, Associate Professor of digital media in the Program in Writing and Humanistic Studies at MIT, has discussed *Mindwheel* in several venues, both in refereed books and conference papers, but also on the Interactive Fiction Wiki, a participatory media project to which he contributes (Montfort, *Twisty* 174-181; "Condemned"). Montfort reaches the same conclusions as Wikipedia about who owns the IP rights to the game, and while he does not footnote the "encyclopedia that anyone can edit," it seems likely that he relied on the extensive network of user-generated game resources that he avidly promotes, uses, authors, and edits. The process by which we attempted to acquire information about the IP rights of Mindwheel thus serves as an object lesson in the way knowledge circulates across professional and amateur communities of practice, and the degree to which these communities loop back on each other.

Moving beyond our personal experiences with the PVW project, we would observe that from the perspective of collection development, the professional archivist and the private archivist often closely resemble each other. Many collections of videogame materials vary in scope, focus, and validity. Some are housed in large academic institutions, within the context of traditional archives; invested amateurs run others, and these collections are limited by the time and resources of that individual or group of individuals. Some of the collections focus on videogame material; others collect videogame material as a sub-set of their main institutional goal. Finally, some of these collections try to provide access to primary materials while others either provide access to end products (like video games, manuals, or individual music tracks), or they collect and organize information about end products. Many of these collections, whether housed within institutional settings or in individual server space, are managed by people who have an intense personal interest in their success; videogames are their avocation rather than their vocation, and the videogame collection isn't necessarily the main collection's focus. Due to this lack of decisive institutional support, even within institutions, videogame collections as a whole appear haphazard and fractured as opposed to collections that focus on conventional subjects, or are composed of traditional materials.

The fact that invested individuals are in charge of these collections can either be a strength or weakness depending on your point of view; for example, invested amateurs have created emulated versions of out-of-print games—these emulated games might be illegal (in that they are still intellectual property of some entity—although "abandonware"), and focus on game play rather than authentic preservation. Whatever the case, these individuals' interest, energy, and devotion are characteristics that will have a guiding influence on the development of formal institutional collections. Additionally, these somewhat ad-hoc collections set the stage for future development of more formal repositories, and provide the information professions with a valuable starting point for future collection development.

References

"Broøderbund." *Wikipedia: The Free Encyclopedia*. October 15, 2008. Retrieved November 12, 2008: http://en.wikipedia.org/wiki/Broderbund

Committee for Film Preservation and Public Access (1993). Preservation without Access is Pointless: Statement by The Committee for Film Preservation and Public Access before The National Film Preservation Board of the Library of Congress, Los Angeles, California, February 12, 1993. Retrieved March 11, 2009: http://www.loc.gov/film/pdfs/fcmtefilmprespubaccess.pdf

Hedstrom, M., Lee, C., Olson, J., & Lampe, C. (2006). "The Old Version Flickers More: Digital Preservation from the User's Perspective." *American Archivist*, 69(1), 159-187. Retrieved August 4, 2008: http://archivists.metapress.com/content/1765364485n41800

Montfort, Nick. “Condemned to Reload It.” *Nick Montfort*. November 11, 2003. Retrieved November 12, 2008: http://nickm.com/writing/essays/condemned_to_reload_it.html

Montfort, Nick. *Twisty little passages : an approach to interactive fiction*. Cambridge, Mass.: MIT Press, 2003.

Preserving Virtual Worlds. November 10, 2008. University of Illinois Urbana-Champagne; University of Maryland, College Park; Stanford University; and Rochester Institute of Technology. Retrieved November 13, 2008: http://pvw.illinois.edu/pvw/

“Writing the history of virtual worlds.” *BBC*. Interview with Megan Winget. August 15, 2008. Retrieved November 12, 2008. <http://news.bbc.co.uk/2/hi/technology/7561553.stm>

Note

[1] On Megan Winget’s Game Preservation Project, see her interview with the BBC cited in the bibliography.

New World Ordering

Chair: Bethany Nowviskie

Bethany Nowviskie
University of Virginia Library Scholars’ Lab
bethany@virginia.edu

Joseph F. Gilbert
University of Virginia Library Scholars’ Lab
joegilbert@virginia.edu

Kelly Johnston
University of Virginia Library Scholars’ Lab
kgj3t@virginia.edu

Christopher Gist
University of Virginia Library Scholars’ Lab
cgist@virginia.edu

Adam Soroka
University of Virginia Library Scholars’ Lab
ajs6f@virginia.edu

In April of 2008, LLC published a thorough survey by Martyn Jessop of many factors contributing to an “inhibition” of the use of GIS, or Geographical Information Science, in the digital humanities (Jessop 2008). That GIS has been slow to penetrate a scholarly population generally receptive to new practices and technologies begs a discussion of issues at once historical and methodological, institutional and pragmatic. It also demands serious engagement by scholars, programmers, librarians, and advocates for shared data and transparent, flexible services. To be effective, this engagement must come at many levels simultaneously: we must work to build core infrastructure to support GIS and leverage the strengths of (primarily government and academic) data providers; we must carefully analyze past successes as well as failures in the digital humanities in order to move forward with more robustly-imagined scholarly projects; and we must interrogate both a toolset that has evolved to suit scientific inquiry (that is, positivist models of physical behavior and dense, detailed, precisely-defined data sets, generally synchronic) and our own inherited systems for interpreting the human record within a spatial field, georeferenced or conceptual. Above all, we must make a concerted effort at understanding what it is we do, when we “do GIS.” This panel will provide differing perspectives on GIS for humanities scholarship, but from within the coherent narrative of a University of Virginia Library effort to build modern infrastructure, support innovative digital projects, and open up dialogue about the causes and conditions of our community's strange inhibition.

UVA Library's Director of Digital Research & Scholarship, Bethany Nowviskie, will open this discussion, but in order to place it within a real-world services framework, Scholars' Lab GIS Specialist Kelly Johnston and programmer Adam Soroka will describe our efforts in building an infrastructure that includes sophisticated hardware and several layers of software: a datastore for standards-compliant metadata, vector, and raster information; a translation layer that provides Open Geospatial Constortium (OGC) services to GIS applications and directly to the Web; and an application layer including a discovery portal and browser-based, interactive mapping systems that consume OGC web services.

For much of its history, GIS work has been predicated on a model of powerful independent workstations featuring complicated monolithic closed-source software, large locally-stored datasets, and users working in relative isolation who publish results in traditional media (Lo and Yeung). Until recently, this has been the model in play at the University of Virginia Library and at most scholarly institutions. A new way of working, introduced by the OGC and gaining increasing traction, utilizes Web services such as Web Map Service (WMS) and Web Feature Service (WFS) and creates an environment where processing occurs server-side, data is stored centrally and shared, and both processes and results can be published digitally, in networked media. The UVA Library Scholars' Lab brought this new way of working to our local teaching and research community through a GIS spatial data infrastructure project (colloquially called the "New World Order"), in continued development since 2008. Its tools enable new forms of collaboration, new forms of publishing, and new forms of pedagogy – and they enforce better internal stewardship of GIS datasets through standardized metadata.

As is common, the UVA Library geospatial data collection has grown episodically, in many formats over many years, in response to user need and data availability. In 2007 we resolved to find a strategic solution to streamline user access to datasets and improve our ability to respond to user requests. Johnston will review the criteria used in our selection process and explain our software choices. He will address: how this new infrastructure better supports collaboration among scholars building spatial datasets across projects and disciplines; how it undergirds the Library's efforts in classroom support by simplifying the contribution and sharing of datasets for collaborative pedagogical activities; and how we identified standard metadata as a key to success and embarked on work to improve both the quantity and quality of our spatial dataset.

To illuminate the advantages of this approach from a developer's standpoint, Scholars' Lab GIS Specialist Christopher Gist will offer a case study in digital scholarship using GIS technologies, new and old. In 2005, UVA Library's GeoStat Center (now part of the Scholars' Lab), in cooperation with the Virginia Center for Digital History, developed a Web-based student collaboration tool for a history class being taught by Professor Ed Ayers. The Southern History Database (SHD) aimed to allow students to pool their research on primary sources to provide larger datasets than would be possible for individual students to collect. The process was also meant to be iterative, so that one class might build on the work of previous semesters.

The SHD had a large spatial component, which involved students' determining the location of events referenced in their primary sources. These results were then tallied and presented in thematic map form. Users could click the map to retrieve results for a given location. This portion of the project presented some very difficult technical issues for the development group. Traditional web mapping applications require specialized server applications and incurred great expense in licensing and management. Available open source applications were limited to static data.

Gist will describe how the development team cobbled together several applications, including MapServer, using the scripting language Perl to create a custom map file on user request. This method is server intensive, difficult to maintain long-term, and not scalable. If new features were to be added, existing scripts must be completely rewritten.

Since that time, many new open source applications have risen from the OGC movement to standardize spatial services and applications, and are available through the Open Source Geospatial Foundation. Since all these tools use common standards – the same ones on which we are building our GIS infrastructure at UVA – it has become easy to develop mapping applications that allow for the creation, storage, management, delivery and display of spatial data.

Now instead of large and expensive proprietary spatial web applications, the paradigm has moved to lighter tools that can be used together through OGC standards. This will allow us nimbly to create complex applications like the SHD in a matter of days instead of weeks or months. And unlike one-off applications, these new applications are flexible and scalable with outputs in many formats that can be used in Google Earth and other easily-acquired tools. All these features make spatial applications for humanities scholarship much more attainable.

But will they help us overcome our essential inhibitions when it comes to GIS in the digital humanities? These locate themselves not only, as Jessop reminds us, in questions about the nature of the tools and our facility with them, but also in more theoretical and methodological concerns surrounding the status of image-based scholarship and our critical stance toward space and place. Bethany Nowviskie and UVA Scholars' Lab coordinator Joe Gilbert continue the discussion with a look at the exigencies of humanities information and our critical impulses.

How do we conceive of and create spatial tools that are neither strictly mimetic nor strictly symbolic? One obstacle we may encounter is that maps, particularly in their digital embodiments, primarily seek to represent real-world spaces and objects. We are therefore limited by these tools' seeming inability to engage objects on multiple semiotic levels. If we can only use spatial tools to investigate what C.S. Peirce termed "iconic," or representative, types of signs, we are artificially restricting the kinds of interpretive acts with which we can engage (Peirce). Current GIS tools elide the subjective nature of both time and space, and thus the possibility of using such tools as richly signifying spaces in themselves.

Nowviskie will offer a young lady's 1823 "Book of Penmanship" from the David Rumsey Historical Map collection as an example of a sophisticated (if naïve) humanities GIS—an historical and artistic document that embeds a brand of subjectivity that the tools and frameworks we build must be able to accommodate (Henshaw). This artifact—with its fascinating imaginative reconstructions, in text and image, of 1820s American geography—will also be presented, more polemically, as the physical embodiment of a method that we must take seriously in digital humanities scholarship if GIS is to take hold: graphesis, the visual construction of knowledge.

Gilbert will conclude with a look at a spatial signifier that has become iconic in a completely different sense – the Google Map marker. Here we encounter another potential difficulty for spatial analysis in the humanities: the symbolic language used by prevalent tools is too divorced (to paraphrase Raymond Carver) from what we talk about when we talk about the world. Here, the fear of maps as representations of the real is avoided, but is also replaced by an empty, literally (or virtually) floating set of signifiers. Rather than depicting this limited sign system laid over the rigid lattice of latitude and longitude, can we re-think Google Earth's virtual sphere of photorealism and arbitrary markers through the lens of Yuri Lotman's concept of the semiosphere, a "synchronic semiotic space?" (Lotman). A future avenue for investigation implicit in this argument would be an understanding of mapping interfaces in terms of Baudrillard's notion of simulacra.

References

Henshaw, Frances A. "Frances A. Henshaw's Book of Penmanship Executed at the Middlebury Female Academy." April 29, 1823. In the collection of David Rumsey. Pub list no. 2501.000. http://davidrumsey.com/

Jessop, Martyn. "The Inhibition of Geographic Information in Digital Humanities Scholarship." Literary and Linguistic Computing, vol. 23 n. 1; April 2008.

Lo, C.P. and Yeung, Albert K.W., Concepts and Techniques of Geographic Information Systems. New Jersey: Pearson Prentice Hall. 2007.

Lotman, Yuri. Universe of the Mind: A Semiotic Theory of Culture. Bloomington: Indiana UP. 2001.

Peirce, C.S. Writings of Charles S. Peirce: a chronological edition. Ed. Fisch, Max H. Bloomington: Indiana UP. 1982.

Use cases driving the tool development in the MONK project

Chair: Catherine Plaisant
University of Maryland, College Park

The Mellon funded MONK project (www.monkproject.org) is developing a digital environment to help humanities scholars discover and analyze patterns. During the past two years of this project major leaps have been made in the development of the infrastructure necessary to 1) normalize and ingest text into a datastore, 2) support data mining and rich text analytics and 3) provide a flexible workbench with a variety of user interfaces for scholars to conduct their analysis tasks. The design and development process of this environment was guided by a small set of representative use cases. While the development of the MONK environment is still ongoing, scholars have been able to use early prototypes as well as other research or off-the shelf tools to conduct their scholarly work, while informing the development team as to the useful features to be incorporated in MONK environment.

In this session we report on three MONK use cases which are interesting individually but collectively illustrate the type of tools that are being developed by Monk to support other scholars in the future. The use cases were selected to be diverse, representative of the questions MONK aim to address, and driven by a scholar already actively engaged in the study of the question. The scholars were selected from the institutions of the project's principal investigators. In the discussion period we hope to discuss the struggles and successes our team encountered in the process.

The rest of this statement summarizes the status of the MONK project.

As of the time of submission MONK includes approximately 1,200 texts, including 300 American novels published between 1851 and 1875, 250 British novels published between 1780 and 1900, 300 plays, 30 works of 16th and 17th century poetry, and some 300 works of 16th and 17th prose, including fiction, sermons, travel literature, and witchcraft texts.

The texts are normalized (using Abbot, a complex XSL stylesheet) to TEI -A, and each text has been "adorned" (using Morphadorner) with tokenization, sentence boundaries, standard spellings, parts of speech and lemmata, before being ingested into a database that provides Java access methods for extracting data for many purposes, including searching for objects; direct presentation in end-user applications as tables, lists, concordances, or visualizations; getting feature counts and frequencies for analysis by data-mining and other analytic procedures; and getting tokenized streams of text for working with n-gram and other collocation analyses, repetition analyses, and corpus query-language pattern-matching operations. Finally, quantitative analytics like naive Bayesian analysis, support vector machines, Dunnings log likelihood, etc., are run through the SEASR environment.

MONK will combine texts and tools to enable literary research through the discovery, exploration, and visualization of patterns. Users start a project with one of the toolsets that has been predefined by the MONK team. Each toolset combines individual tools (e.g. a search tool, a browsing tool, a rating tool, and a visualization) that are applied to worksets of texts selected by the user from the MONK datastore. Worksets and results can be saved for later use or modification, and results can be exported in some standard formats (e.g., CSV files).

Paper 1: Formulaic Emotion: Reading Victorian Deathbed Scenes from a Distance

Sara Steger
University of Georgia
ssteger@uga.edu

There is perhaps no other scene as quintessentially Victorian as the deathbed scene. In one of his more humorous observations, Garrett Stewart observes: "Some characters must die in any period of novel writing. As everyone allows, characters die more often, more slowly, and more vocally in the Victorian age than ever before or since" (8). A scene such as the death of Little Nell in *The Old Curiosity Shop*, which has now come to represent over-the-top Victorian sentimentality, once "sent most of literate England into mourning" (Kuchich 59).

In deathbed scenes, Victorian authors convey much more than just an emotional moment that advances the plot. In the same manner in which Victorian mourning rituals became formal indicators of the status and worthiness of a deceased person, authors adopted patterns serving as markers to the reader about how to feel about the deceased character. Like mourning practices, such fictionalized scenes were expected to fit a formula, designed to generate sympathetic feelings in the reader. Deathbed scenes solicit an affective response by meeting the readers' expectations for a "good death" and enveloping these scenes with a certain bittersweet pathos. The formal and methodical practices of Victorian mourning thus manifested in deathbed scenes of the period. In these deathbed scenes, there is a code within the language that aims to elicit an affective response in the reader.

In my work[i], I use a combination of close and distant reading lenses to uncover the low-level patterns in vocabulary that equate to the higher-level formation of deathbed scenes as a literary topos, a framework within which authors can utilize certain linguistic and thematic patterns to manipulate the emotions of readers. I am building upon Franco Moretti's conceptions of a "quantitative approach to literature" (4), which I find productively compatible with Garrett Stewart's structuralist interpretation regarding the "style of the death sentence," in which he argues that death scenes use specialized rhetoric (figures of speech and grammatical devices), which "can be sorted out if not strictly codified" (7).

I begin by employing WordHoard, a "philological tool" designed to precipitate "the close reading and scholarly analysis of deeply tagged texts" (http://wordhoard.northwestern.edu) to build lexicons of my workset of sixteen deathbed scenes. These lists form the beginning of the ways that micro-patterns parallel larger thematic patterns in the texts. For example, the Victorian ideal of the "good death" is apparent in the vocabulary: "good," "better," and "great" all appear in the top twenty adjectives. There is a physical materiality in the list of nouns, which includes specific parts of the body: the "hand," the "face," the "heart," and the "eye." Finally, in both the list of nouns and the lists of verb, words that reflect an interest in last looks and words stand out: "word," "speak," "tell," "look," "see," "say," and "hear."

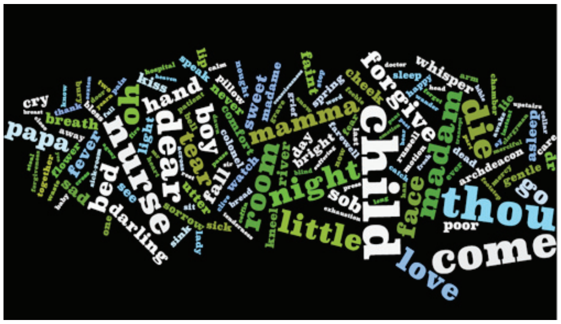


Figure A. Words that are Over-Represented in Victorian Deathbed Scenes Compared to Testbed

The raw count of a word is not always the best measure of its significance, however, so I also employ Dunning's Log Likelihood Ratio to provide a measurement of which words were over-represented and which were under-represented in the deathbed scenes compared to my testbed set of eighty mid-Victorian novels. I thus was able to obtain a statistical measure of words that are significant in deathbed scenes. I then input the full list of significant words into a Wordle wordcloud visualization (see Fig A).[ii] By using Dunning's log likelihood ratio, I was able to produce visualizations not just of the most common words, but of the words that are most salient in deathbed scenes.

The words that are over-represented in deathbed scenes correspond to the thematic patterns of deathbed scenes identified above. Words that set the scene are scattered across the visualization, hinting at descriptions of the bed, the room, the pillow, the chamber, and even the hospital. Words corresponding to illness also are prominent, including "fever," "sick," "nurse," "doctor," and "sick-room." Moreover, the word cloud is a reminder of how death is a domestic affair. Not only is there an emphasis on that most domestic of spaces, the bedroom, but the vocabulary emphasizes intimate relationships—"mamma," "papa," "darling," and "child." This latter word also crosses over into demonstrating a concern with innocence and diminutiveness, especially when read with to the related "baby" and "little." The visualization also reflects the thematic importance of last words and touches—the "lips" that "speak," "whisper," or "kiss," the "last" "farewell," the final "breath." Emotions run high in these scenes, which are filled with "sob[s]," "tear[s]," and "cry[ing]." The language captured in the word cloud also appears somewhat elevated, with the prominent "thou," the lofty and interestingly negative "nought," and the somewhat perplexing presences of both "madame" and "madam." Finally, we again see evidence here in the vocabulary of the deathbed scene of the idea of a "good death"—the loved ones are sent to "heaven" having obtained "forgiveness" or having learned to "forgive." They have earned "mercy" and are "happy" in their final moments.

While a close reader may be able to get a sense of which words are used more often in sentimental scenes, the Dunning's Log Likelihood ratio enabled me to discover information that a scholar would never be able to obtain without these technologies: that which is absent. What the word cloud does not include is almost as informative as what it does. Given the prominence of mourning in Victorian culture, there is almost no trace of the formal trappings of mourning in this snapshot of deathbed scenes. While the words "coffin," "archdeacon," and "grave" appear, the visualization shows that the topos is much more concerned with describing the death than with detailing the mourning. The authors, it seems, rely less on the pomp and ritual of ceremony to convey a sense of the character's worth. A description of the moment is sufficient to convey the "good death;" the burial and the mourning – the public moments in the church and graveyard – are largely absent. While this seems to contradict my earlier hypothesis that deathbed scenes are reflections of Victorian mourning practices, the act of writing itself creates the public moment of grief otherwise missing from the narrative. Authors brought the intimate moment of death to the public eye, effectively drawing the reader and creating a community of mourners. It is the relationship between the reader and the character that serves as proof of the character's "worth."

This makes the list and visualization of the words that are under-represented in deathbed scenes even more striking (see Fig B). One of the most under-represented words is "holy," and it is followed by "church," "saint," "faith," "believe" and "truth." It seems the Victorian deathbed scene is more concerned with relationships, marked by words such as "forgiveness," "mercy," "forgive," and "comfort" than with personal convictions and declarations of faith. The deathbed scene is a moment in which the dying person connects one final time with living. As such, there is an emphasis on embodiment in the over-represented words: "cheek," "breast," "hand," and "face." These moments are presented less as "holy" than they are as deeply, profoundly, human moments.

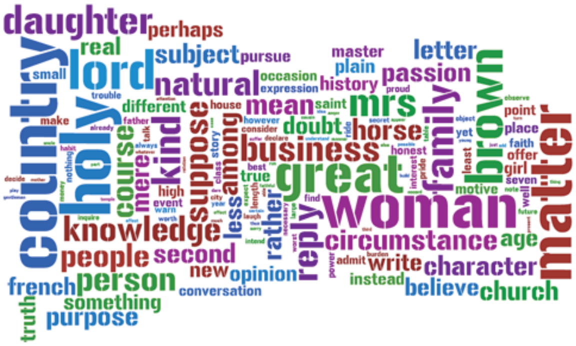


Figure B. Words that are Under-Represented in Victorian Deathbed Scenes Compared to Testbed

Besides the surprising lack of emphasis on holiness and faith, the word cloud reveals and highlights other words that don't represent the essence of the Victorian deathbed scene. Words that have to do with business and class ("money," "power," "business," "lord," and "gentleman") don't belong at the deathbed. There's also a trend toward generalities, including general groups of "people": "person," "woman," and "girl." Even though "family," "daughter," and "father" make the list of under-represented words, they are more generic terms when compared to the more familiar "mammas" and "papas." Tellingly, words of uncertainty also appear prominent in the visualization of under-used words in deathbed scenes, including "suppose," "perhaps," and "doubt." The deathbed scene, by nature a scene of resolution, leaves no room for incertitude. Altogether, the words that are not used, the "negatives," serve as a sort of shadow to the "positives," giving dimension to the themes and patterns that stood out in the first visualization.

The increased scope, scale and speed of text analysis take me beyond my preconceived notions about the texts, revealing trends visible only with the perspective of distance. Rather than representing the end result of the experiment, the data from my experiments inspired more sharply focused readings of the texts. A visualization like the Wordle image of under- and over-represented words should not, and really cannot, stand as evidence in proving a hypothesis. The visualization is simply not empirical in nature. In a way, word clouds, as visual representations of criticism, can be seen as art useful in representing other modes of art. And, as with any instance of artistic representation, they remain open to interpretation. Distant reading can only be a viable methodology for literary study if it is used in this way: as a new lens for examining the text and as a means to inspire interpretations and readings based on the new perspective.

Notes

[i] The project is also one of the "use cases" that drives development for the MONK Project (http://monkproject.org).

[ii] Please see http://wordle.net

References

Kuchich, John. "Death Worship among Victorians: The Old Curiosity Shop." PMLA 95.1 (Jan 1980): 58-72.

Moretti, Franco. Graphs, Maps, Trees: Abstract Models for a Literary Theory. New York: Verso, 2005.

Stewart, Garrett. Death Sentences: Styles of Dying in Victorian Fiction. Cambridge, Mass: Harvard UP: 1984.

WordHoard. Northwestern University, 2008. <http://wordhoard.northwestern.edu>.

Wordle. Jonathan Feinberg, dev. <http://www.wordle.net>.

Paper 2: The Devil and Mother Shipton: Serendipitous Associations and the MONK Project

Kirsten C. Uszkalo
Simon Fraser University
kcu2@sfu.ca /kirsten@uszkalo.com

Datamining tools such as Naïve Bayes, which seek to extract and quantify textual features, do not appear to support pedagogies, research practices, or facilitate the pleasures of discovery in the same way humanist researchers have become used to (Rommel 2004, Harley et al. 2006). Jean Guy Meunier, Ismail Biskri, and Dominic Forrest (2005) argue that reading or analyzing text is "a more complex procedure of heuristics and pattern recognition strategies. All of which are grounded on complex dimensions such as linguistic and mundane semantic structure, inference, pragmatic memory, culture, social interaction, and knowledge repositories" (124, 126). They conclude the "text does not necessarily reveal itself in a first reading, not even in a first analysis" and that the "computer can play a productive role in the reading and analysis process; but only if it is well-situated as an assistant to such cognitive activities" (126). Although in its current stage, the MONK Workbench may be most helpful to domain experts who are already familiar with the social, legal, or literary phenomenon they study, it can be used to facilitate pattern finding across corpora large enough to loosen the threads that tie texts together. My original research question asked if analytics used within the MONK Workbench could help locate Richard Head's biography of Mother Shipton, not within the fifteenth-century, when he claims it was originally written, but within the late sixteenth century when it was published. While reporting on these preliminary findings, this paper will likewise suggest that, although the tools can render texts into composite features, they can also suggest ways in which texts can be meaningfully realigned. The serendipitous associations created by these re-alignments support an engagement with texts that is familiar to research scholars working outside of digital humanist practices.

Mother Shipton and the Devil

Mother Shipton may have lived in the late 15th to mid 16th centuries (1488-1561?); she emerged into the literary record in 1641 as an early modern prophetic witch whose embodied spirituality was made her spiritual allegiances hard to identify. Her prophecies were ultimately created or appropriated by authors who tantalizingly propose that her predictive skill and bodily deformities illustrated an otherworldly or demonic heritage. The inclusion of the Gentleman Devil, a sucked witch's mark, and animal familiars in Richard Head's biographies of Mother Shipton situate these texts alongside the prophetic norms and evolution of theories of witchcraft and diabolism written during the seventeenth century. In hoping to exploit interest in ecstatic prophecy, practical maleficium, and, most crucially, the Devil's relationship to the witch, Head's biographies conform too closely to seventeenth-century stylistic elements, ultimately revealing their fictions. The devil is in the details.

I theorize that the presence of this seducing gentleman devil could help situate Head's construction of Mother Shipton within a context of continental beliefs about witchcraft emerging in seventeenth-century trial accounts. In England, the witch's mark appeared most often as a scratch, or a prick, then as a teat from which the witch's familiar was supposed to drink as reward for the malefic crimes it committed at the witch's behest. References to having sex with the devil became prominent in England as a reflection of newly imported continental beliefs about witchcraft. The devil in the shape of a gentleman, or the "black man," who asks to suck from these increasingly sexualized marks, appears after the sixteenth-century when Head's biography of Shitpon was supposedly written. At the beginning of my MONK research I created a paper dataset with pens and highlighters. With this research, I found a tract, "The Mystery of Witchcraft" (1626), which showed the presence of "carnall knowledge of [the witches] body" earlier than I'd anticipated. In finding this, I had to rethink when the idea of demonic sex entered the understanding of English witchcraft, but also whether coitus was the only demonic sex act present in the material.

The MONK Workbench combines "texts and tools to enable literary research through the discovery, exploration, and visualization of patterns. . . . Each toolset is made up of individual tools (e.g. a search tool, a browsing tool, a rating tool, and a visualization), and these tools are applied to worksets of texts selected by the user from the MONK datastore" (MONK, Workbench). I rated a large sample set of fifty-two texts using the Search by Example tool, and rated them as containing "sex," "no sex," and "some sex" with the devil, and ran a Naïve Bayes analytical routine, asking for lemma as my feature, and requesting one hundred "features" (words that are representative of each class of documents) be returned. Some interesting results emerged. For example, the later witchcraft tract, *A True and impartial relation of the informations against three witches, viz., Temperance Lloyd, Mary Trembles, and Susanna Edwards, who were indicted, arraigned and convicted at the assizes holden for the county of Devon, at the castle of Exon, Aug. 14,* *1682* was returned as a match for "sex." Here is a suggestive excerpt found in the tract:

> The said Informant upon his Oath saith, That upon the 17th day of July instant this Informant did hear Susanna Edwards to confess, that the Devil had carnal knowledge of her Body; and that he had suckt her in her Breast and in her Secret parts.

The Workbench analytics returned *From Newes from Scotland, Declaring the Damnable life and death of Doctor Fian, a notable Sorcerer, who was burned at Edenbrough in Ianuary last. 1591* as a positive hit for "sex" with the devil. This excerpt, extracted from a large enough text chunk to make it difficult to locate by eye, comments, perhaps unintentionally, on the more pleasurable kinds of malefic sexuality seen in these encounters: for as much as by due examination of witchcraft and witches in Scotland, it hath latelye beene found that the Deuill dooth generallye marke them with a priuie marke, by reason the Witches haue confessed themselues, that the Diuell dooth lick them with his tung in some priuy part of their bodie, before hee dooth receiue them to be his seruants, which marke commonly is giuen them vnder the haire in some part of their bodye, wherby it may not easily be found out or seene, although they be searched

Although these texts do not share significant linguistic features, and the second excerpt is not in fact a story of English witchcraft, but of Scottish, this represents a very positive result in terms of finding patterns of malefic sexuality across a corpus. Although they suggest the necessity of rethinking how sexuality functions in witchcraft tracts and as such proves to be a useful to look at familiar texts, their location within large text chunks would have obscured them to non-domain experts.

The most provocative results of the sample was a serendipitous association that contextualized Richard Head's construction of Mother Shipton's diabolical upbringing. The MONK Workbench provided an opportunity to "search by example" and find unpredicted associations in terms of this research stream. The Workbench analytics rated a text as "sex" with the devil, which I had not considered, a text about Tannakin Skinker, the "hog-faced woman" from Holland. Head's text and the Skinker text share themes such as implied witchcraft, the presence of the Devil, and the implication of the Devil's presence, the presence of a witch or a witch's curse, the use of wealth as a lure, the use of costly clothing as a tool for seduction, the presence of a castle or a large home, and, most substantially, hog-faced children. These numerous alignments are suggestive of the kinds of source text or framework upon which Head drew an expanded version of Shipton's life. Although malefic sex was identified as a signifier of diabolism, numerous texts rated as "sex" with the devil by the Workbench analytic process included prodigious birth as a sign of the Devil's presence. Prodigious births, along with witchcraft, comets, and two-headed chickens, belongs to the same kind of tabloid, or "enquirer," genre in early modern England; Head's version of Shipton's life falls firmly within that genre in a way which makes the devil's presence in it make sense outside of witchcraft or prophecy.

Conclusions

In this way, the computational process assisted my reading process rather than delineated it as Meuneir et al. argue must happen to make computational analytics more useful to main stream critical inquiry. Although Head's version of Shipton's life is not a text in the dataset, I was able to use the MONK Workbench to find a set of texts within the same genre. That is, I was able to find comparable texts without a base text against which they were compared. In addition, while the "predicted features" were only moderately helpful, patterns began to emerge that reified and challenged my understanding of the devil's evolution in sixteenth-century witchcraft accounts and therefore my understanding of Mother Shipton's location within these texts. Although preliminary trials with Workbench analytics produced mixed results, only sometimes predicting that texts contained the elements I was looking for to help contextualize Richard Head's version of Mother Shipton within sixteenth century ephemera, it suggested connections based on similarities and patterns I had not predicted. As a result, I reevaluated my hypothesis regarding demonic sexuality and its role in the prodigious birth category of early English ephemeral publications. Ultimately, this experience facilitated serendipitous literary discoveries based on unpredictable features, a process of discovery that makes tools such as those incorporated in the MONK Workbench more user-friendly to users unfamiliar with current digital research methods.

Works cited

Anon. A True and impartial relation of the informations against three witches, viz., Temperance Lloyd, Mary Trembles, and Susanna Edwards, who were indicted, arraigned and convicted at the assizes holden for the county of Devon, at the castle of Exon, Aug. 14, 1682

Anon. The most strange and admirable discouerie of the three witches of Warboys. London: Printed for Thomas Man and Iohn Winnington, and are to be solde in Pater noster Row, at the signe of the Talbot. 1593.

Anon. The prophesie of Mother Shipton in the raigne of King Henry the Eighth. London: Printed for Richard Lownds, at his Shop adjoyning to Ludgate, 1641.

Anon. The Strange and wonderful history of Mother Shipton. London: Printed for W.H. and sold by J. Conyers, 1667/1668?

Head, Richard. The life and death of Mother Shipton, London : Printed for B. Harris 1677.

Harley, Diane, et al. Use and Users of Digital Resources: A Focus on Undergraduate Education in the Humanities and Social Sciences (2006) Center for Studies in Higher Education, University of California, Berkeley. 11-06

Meunier, Jean Guy, Ismail Briski, and Dominic Forest. "A Model for Computer Analysis and Reading of Text (CARAT): The SATIM Approach" Text Technology, Number 2, 2005. 123-151

MONK Project. "Introduction" Monk Workbench. Online.

Rommel, Thomas. "Literary Studies." A Companion to Digital Humanities.

SCHREIBMAN, SUSAN, RAY SIEMENS and JOHN UNSWORTH (eds). Blackwell Publishing, 2004. Blackwell Reference Online. 22 September 2008 http://www.blackwellreference.com.login.ezproxy.library.ualberta.ca/subscriber/tocnode?id=g9781405103213_chunk_g978140510321311

Paper 3: The Story of One: Developing text mining procedures and visualizations in the MONK project to re-read Gertrude Stein's The Making of Americans.

Tanya Clement
University of Maryland, College Park
tclement@umd.edu

Catherine Plaisant
University of Maryland, College Park
plaisant@umd.edu

Romain Vuillemot
HCIL Human-Computer Interaction Lab

Most critiques of *The Making of Americas* (Paris 1925) by Gertrude Stein contend that the text deconstructs the role narrative plays in determining identity by using indeterminacy to challenge readerly subjectivity. The current perception of *Making* as a postmodern text relies on the notion that there is a tension created by frustrated expectations that result from the text's progressive disbandment of story and plot as the narrative unweaves into seemingly chaotic, meaningless rounds of repetitive words and phrases. Yet, a new perspective that is facilitated by digital tools and based on the highly structured nature of the text suggests that these instabilities can be resolved by the same seemingly nonsensical, non-narrative structures. Seeing the manner in which the structure of the text makes meaning *in conversation with* narrative alleviates perceived instabilities in the discourse. The discourse about identity formation is engaged—not dissolved in indeterminacy—to the extent that the reader can read the composition.

One method for reading the composition of the text without relying on what becomes a non-existent framework based on plot is to view the progression of words according to a different framework, a framework that relies on comparative associations based on word usage. Using *WordHoard*[i], we compared word usage between texts and text parts by calculating the log-likelihood ratio, which describes the size and significance of the difference between word frequencies in a base text versus a reference text.[ii] In this analysis, we measured word usage in The Making of Americans in comparison to two different sets of reference texts with more traditional narrative structures: (a) a set of 19th century novels written by Jane Austen, Charles Dickens, George Eliot, and George Meredith;[iii] (b) between the first and second half of Making, which it has been argued also represents different narrative trends (Clement 2008). Visualizing this information in *Wordle*[iv]—a word cloud application (Wattenberg & Viegas, 2008)—is useful primarily because it provides a visual overview of word frequencies that is easy to understand and to publish for reference. The *Wordle* application facilitates this kind of analysis by visualizing the list of words in a cloud that maximizes the space utilization on a computer screen by sizing the words by their relative frequencies. The more frequently a word occurs in a particular text (relative to another text) the larger the word appears. In the set of visualizations that accompany this discussion,[v] each cloud serves to visualize words that are more or less frequent in any given comparison (see examples in Fig. 1). What becomes immediately evident in comparing these visualizations is the prominence of a particular word that consistently scores a high value in terms of discrepancy between *Making* and the reference texts: *one*.

The word *one* appears consistently across every cloud that marks the words that are more common in *Making* and less common in the sample of nineteenth century texts and words that are more common in the second half rather than the first half of the text. We then compared the relative frequencies of multiple pronouns, which revealed that the frequency of *one* surges by the end of the text (Fig. 2). What is most interesting about this graph is that the high frequency of one is the result of the confusion accomplished by the word's schizophrenic nature. Words here are represented according to occurrence, not to type of occurrence. Thus, the word *one*—unlike *he*, *she*, *I*, *we*, or even *you* or *it*—which plays many positions in the text, in the role of a pronoun or an adjective and in the subject or object position, surges in frequency.

To better understand how the word one is used in the text, we created another set of visualizations prepared using a prototype we developed called *PosViz*.[vi] *PosViz* allowed us to compare word usage based on parts of speech in individual chapters from Making to the whole text. These comparisons allowed us to isolate and analyze words used more and less frequently throughout the course of the novel and to measure how and if these patterns change within the text itself. In these visualizations, each part of speech for each word is treated as a separate word instance and each instance is placed according to its appearance in the text, color-coded according to its part of speech, and sized according to its overall frequency. Thus, by using *PosViz*, the progression of the manner in which the word *one* is used in terms of different parts of speech is documented, allowing us to see that the use of *one* appears to change as the text progresses. For example, a relatively small one appears three times in the chapter 1 cloud (Fig. 3). This visualization indicates that the occurrence of the word one has little variance in terms of how it is used (its part of speech) and occurs relatively infrequently in occurrences that are localized to the beginning paragraphs of the chapter.[vii] By chapter 9, however, *one* dominates the discourse both in terms of its frequency and in terms of its multiple uses (Fig. 4).

By identifying the manner in which word usage changes in correlation to the presence and absence of narrative both in comparison to other novels and within the text itself, these comparisons enable a new perspective on the meaning-making processes of the text's composition. For example, these visualizations illustrate the nature of the word *one* as it is used to heighten the word's propensity for different reading possibilities. This lends to a reading in which one may represent a singular subject position or multiple subject positions at once. With this information, a further argument can be made that the discourse about identity formation is engaged in this multiplicity, not dissolved in indeterminacy. Thus, employing composition in her representation of identity formation in *The Making of Americans* becomes the method by which Stein seeks to represent identity, but if and how the reader is able to recognize and interpret this endeavor is predicated by her ability to see it.

This work with *The Making of Americans* is part of research and development within the MONK (Metadata Offer New Knowledge) project, a Mellon-funded collaborative seeking to develop text mining and visualization software in order to explore patterns across large-scale text collections. Stein's text was a productive text for analysis during the beginning phases of the MONK project since its many and complicated repetitions could be processed and visualized.[viii] This presentation focuses on how the process of determining decision criteria for text mining led to the discovery that various textual features (n-grams, parts-of-speech, and log-likelihood ratios) and various visualizations (*FeatureLens*, *Spotfire*, *Wordle*, and *PosViz*) ultimately facilitated an iterative discovery process and a new reading of Gertrude Stein's *The Making of Americans*.

Appendix of Images

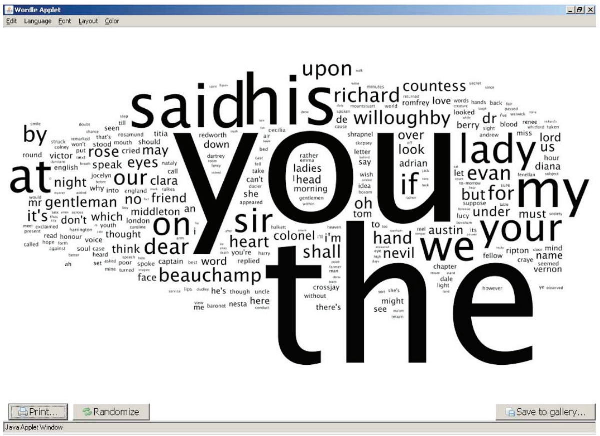


Figure 1: Compasrions between Making and novels by George Meredith, using log-likelihood ratios from WordHoard and visualizations produced in Wordle; A: words that are more common in Meredith;

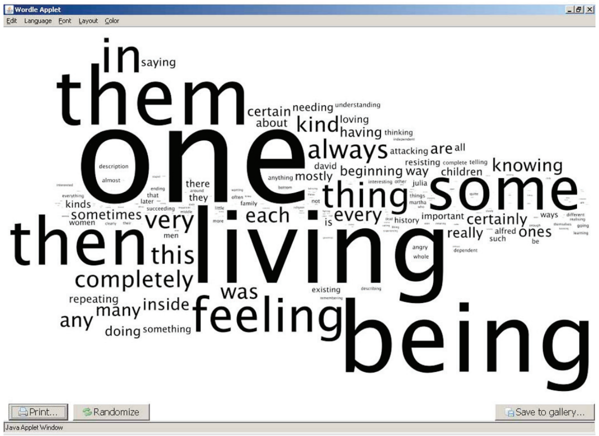


Figure 1.B: words that are more common in Making.

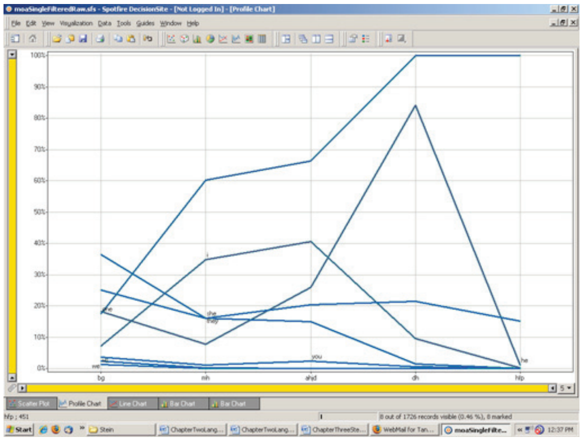


Figure 2. The frequency of "one," "I," "you," "we," "he," "she," and "they" are mapped across the five sections of the text in Spotfire

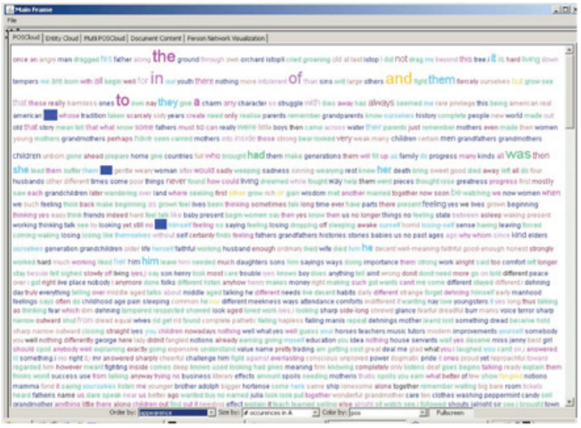


Figure 3: Chapter 1, the word "one" highlighted in PosViz visualization

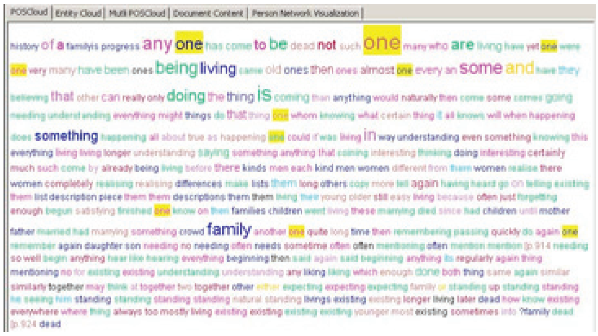


Figure 4: Chapter 9, the word "one" highlighted in PosViz visualization

Notes

[i] Please see http://wordhoard.northwestern.edu/userman/.

[ii] This analysis is based on Dunning's log-likelihood analysis. Please see http://wordhoard.northwestern.edu/userman/analysis-comparewords.html#loglike.

[iii] The books used in this study are those available in the WordHoard application. They are listed at http://terpconnect.umd.edu/~tclement/less-more.htm. These authors were chosen because Stein repeatedly compares her text to their novels. See Stein 1990, p. 506.

[iv] Please see http://wordle.net/.

[v] These visualizations are pictured in an online appendix entitled "Visual Comparisons of Gertrude Stein's The Making of Americans using WordHoard, Wordle, and PosViz" at http://terpconnect.umd.edu/~tclement/less-more.htm. In these slides, the comparisons have been visualized with four sets of data for each data set, all of which are set in comparison to the base text The Making of Americans and include and exclude 'common words' such as articles, conjunctions, and pronouns. The full list of these 'stop-words' is unavailable, but the creator Jonathan Feinberg has indicated in an email that 'I have modified them by hand over time. The English one came from the Snowball stemmer project' at http://snowball.tartarus.org/algorithms/english/stop.txt (personal correspondence).

[vi] Currently, PosViz does not have a web presence.

[vii] The analysis program used to label these uses is part of the SEASR (Software Environment for the Advancement of Scholarly Research) analytic routines (http://seasr.org/). Though imperfect, the system is consistent—it labels the same behaviors the same way each time. Thus each occurrence represents the perception of a different use of the word one.

[viii] This work is published in two articles: Don et al., 2007 and Clement, 2008.

References

Clement, T. (2008). "'A thing not beginning or ending': Using Digital Tools to Distant-Read Gertrude Stein's *The Making of Americans*." *Literary and Linguistic Computing*, 23.3: 361-382.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., Plaisant, C. (2007). "Discovering interesting usage patterns in text collections: integrating text mining with visualization." *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 213-222.

Stein, G. (1990). "Transatlantic Interview 1946." In *The Gender of Modernism*. Bonnie Kime Scott and Mary Lynn Broe (eds). Bloomington: Indiana University Press, pp. 502-516.

Wattenberg, M. and Viegas, F. (2008). "Tag clouds and the case for vernacular visualization." *Interactions*, 15.4: 49-52.

Stein, G. (1995). *The Making of Americans: Being a History of a Family's Progress*. Normal, IL: Dalkey Archive Press.

Weiss, Sholom M. et al. (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer, pp. 85-86.

The Digital Humanities Observatory: Building a National Collaboratory

Chair: Susan Schreibman

Susan Schreibman
Digital Humanities Observatory
susan.schreibman@gmail.com

Jennifer Edmond
Trinity College Dublin
edmondj@tcd.ie

Dot Porter
Digital Humanities Observatory
dot.porter@gmail.com

Shawn Day
Digital Humanities Observatory
s.day@ria.ie

Don Gourley
Digital Humanities Observatory
d.gourley@dho.ie

Introduction

The Digital Humanities Observatory (DHO) was established to support digital humanities research in Ireland, and to manage and coordinate the increasingly complex e-resources created in the arts and humanities throughout the island. The DHO, founded in 2008, has in its first year begun to implement a plan of support focusing on three main issues: encouraging collaboration; providing for the management, access, and preservation of project data; and promulgating shared standards and technology for project development.

The DHO sits solidly in the family of recent international initiatives seeking collaboration, sharing, and preservation, that signal a shift in perspective in the digital humanities environment from a project based (digital silo) approach to one in which the scholarly resources we create are linked, interoperable, reusable, and preserved. Collectively, we have entered a new phase of human and technical infrastructure development. An overview of just a few events from the past few years will serve to put the establishment of the DHO in perspective.

In April 2007 The National Endowment for the Humanities sponsored a meeting gathering some 60 representatives from US digital humanities centres and institutes, and funding agencies that support their work, to discuss short- and long-term priorities and to encourage collaborative opportunities. The result of this meeting was centerNet, "an international network of digital humanities centers formed for cooperative and collaborative action that will benefit digital humanities and allied fields in general and centers as humanities cyberinfrastructure in particular."

In February 2008 the Mellon Foundation-backed initiative, Project Bamboo, began a series of workshops to lay the groundwork for the building of an international infrastructure (technical, social, and institutional) to tackle the question, "How can we advance arts and humanities research through the development of shared technology services?" Unlike the NEH-sponsored meeting that led to the development of centerNet, which drew attendees solely from established digital humanities centers and institutes, the Bamboo Project participants are pulled additionally from humanities departments, IT support, libraries, and administration from each attending organization. These groups are not the traditional practitioners of digital humanities, and Project Bamboo represents a distinct opening up or "popularization" of the field within academia. (http://projectbamboo.uchicago.edu/)

At the same time Project Bamboo was launched, the EU-funded *Interedition: An Interoperable Supranational Infrastructure for Digital Editions* (funded as a COST Action February 2008-April 2012) was launched "to promote the interoperability of the tools and methodology … for digital scholarly editing and research." Interedition and projects like it have the potential to serve both established digital humanities practitioners, such as those represented by centerNet, and those new to the field, such as participants in Project Bamboo. (http://www.interedition.eu/)

Slightly later in the year, in September 2008, the Council on Library and Information Resources published a white paper, "No Brief Candle: Reconceiving Research Libraries for the 21st Century," which sets forth recommendations for collaboration between and among faculty, librarians, and IT professionals.

Conceived in 2004 and with the official kick-off meeting in October 2008, DARIAH (Digital Research Infrastructure for the Arts and Humanities), a network of fourteen partners in ten European countries, is working to develop infrastructure to support the preservation of cultural heritage in Europe and improve access to research material for the humanities. Currently in a preparation phase, the project will begin in 2010. (http://www.dariah.edu/)

Despite the momentum in international collaborative ventures, only a month prior to the April 2007 meeting, the UK's Arts and Humanities Research Council withdrew funding for the Arts and Humanities Data Service (AHDS) from March 2008, despite the fact that the 12 year old network of digital humanities centres of expertise was one of the oldest and most well-respected national infrastructures in the world (press release: http://www.ahrc.ac.uk/News/Latest/Pages/AHRCreshapesitsfundingofICTresearch.aspx).

The withdrawal of funding from the AHDS came just as researchers in Ireland were in the final phases of conceiving the DHO. While those writing the grant realized that they did not want to foster the model prevalent in the early to mid-1990s of the when digital humanities was coalescing as a discipline, i.e. the lone scholar model modified to encompass a small team of postgraduates and technical staff, frequently employed by a digital humanities centre. A model moe in keeping with other international initiatives previously cited was needed: one that reaches across national boarders to encourage shared infrastructures, frameworks, and ontologies. Thus the Digital Humanities Observatory, a national digital humanities centre based in Dublin, Ireland, was founded in 2008 as a response to international developments and to a national need for digital humanities infrastructure in Ireland.

The Digital Humanities Observatory

The DHO was created as part of a larger national infrastructure entitled *Humanities Servicing Irish Society* (HSIS). HSIS is comprised of five of the six Universities in the Republic of Ireland (National University of Ireland, Galway; National University of Ireland, Maynooth; Trinity College Dublin; University College Cork; University College Dublin), the two Universities in the North of Ireland (Queens University, Belfast and University of Ulster), and several institutions of higher education in the Republic (College of Art and Design, Dundalk Institute of Technology, St Patrick's Teacher Training College, Royal Irish Academy).

These institutions came together as a result of a funding call from the Higher Education Authority under the Programme of Research in Third Level Institutions (PRTLI). HSIS was awarded □28,000,000 in August 2007 to build a joint national platform for the coordination and dissemination of humanities research, teaching and training at an all-island level. This is probably the single largest award to the humanities in the world to date. Of that □28,000,000, some □18,000-20,0000 is for capital funding. A majority of the remainder, however, is being invested in digital humanities projects, training, and development.

The DHO is the centrepiece of the HSIS collaborative. It was founded to be a collaborator on national digital humanities initiatives, a centre of excellence, forward looking but cognisant of past humanities and digital humanities practice, and positioned to become a player in international initiatives.

These extremely ambitious goals are being carried out against a backdrop of changing expectations in the roles of digital humanities centres; changing expectations about the resources we create and the nature and rewards for those intellectual products; and a realization that their long-term viability and reusability must be designed for in the very earliest stages of project conception.

To fulfil these goals, the DHO is developing three distinct but integrated technical infrastructures. The first of these is a Portal, which is the public face of the DHO. Built on the content management system, Drupal, it not only provides information about the DHO, its activities and its partners, but features community spaces allowing the Irish academic community, as well as those interested in the work produced by that community, to stay connected.

The second deliverable is a database entitled *Digital Research and Projects in Ireland* (DRAPIer). This projects and methods database (modeled on the UK's ICT Methods Database), also implemented in Drupal, provides a publicly-accessible framework for the discovery of digital humanities project in Ireland. Administratively, DRAPIer has equally important functionally. It allows us to identify projects at risk and to intervene before the content is lost to the academic community. It provides a national snapshot of the depth and breath of digital humanities research in Ireland, funding sources, and methods utilized.

The third deliverable is an access and preservation repository based on Fedora. This repository will provide public access to digital humanities resources created by the HSIS partners. This infrastructure, currently under development, is possibly the most ambitious IT deliverable of the initiative. Some of the resources created by HSIS partners will reside directly in the DHO's Fedora instance, others will be federated in other Fedora instances maintained by DHO project partners. The interoperability of these frameworks is based on shared content modeling within Fedora.

By creating resources in which the underlying data share some level of interoperability, based on common content modeling, shared ontologies, named authority lists, and metadata standards, the DHO expects to provide a level access to a variety of heterogeneous resources having both an eye to the future in their long-term preservation, as well as an eye to the past, in providing access a wealth of Irish cultural heritage, and well as the research of Irish scholars, to a wider audience.

Making primary resources of Irish studies, as well as the research of Irish scholarship more widely available is particularly important for disciplinary studies areas typically labeled as 'minority'. It can be extremely difficult for postgraduate students outside of Ireland or the handful of Irish studies centres outside Ireland to obtain access to the materials they need, further discouraging research in these areas. Digital publication has the ability to change this, leveling the playing field between area studies and more resourced areas such as British or American Literature.

Our panel will consist of four presentations on the political, cultural and technical aspects of the foundation and work of the Digital Humanities Observatory. We envision this presentation being of value to other countries and regional areas wishing to implement a similar multi-institutional centre. Edmond is on the DHO Consultative Committee and was one of the authors of the PRTLI proposal. She will describe the needs behind the DHO and the thoughts behind its initial foundations. Schreibman will introduce the DHO as it is today, and describe its place in the political and cultural landscape of Ireland and introduce some of the projects that have benefited from its support. Porter will discuss the development of the controlled vocabulary that supports DRAPIer and the standards that support the access and preservation repository. Day will demonstrate the Portal and DRAPIer, and will discuss how these deliverables in particular serve the growing community of digital humanities scholars in Ireland.

References

Digital Humanities Centers Summit: Notes from the NEH-hosted summit meeting of digital humanities centers and funders, April 12-13, 2007<https://apps.lis.uiuc.edu/wiki/display/DHC/Digital+Humanities+Centers+Summit>Accessed 14 November 2008

centerNet <http://www.digitalhumanities.org/centernet/>

Arts and Humanities Data Service (AHDS) <http://www.ahrc.ac.uk/News/Latest/Pages/AHRCreshapesitsfundingofICTresearch.aspx>

Project Bamboo <http://projectbamboo.uchicago.edu/>

Interedition: An Interoperable Supranational Infrastructure for Digital Editions <http://www.interedition.eu/>

Council on Library and Information Resources, "No Brief Candle: Reconceiving Research Libraries for the 21st Century," August 2008 <http://www.clir.org/pubs/reports/pub142/contents.html>

Scholarly Communication Institute 6: Humanities Research Centers. University of Virginia, July 13-15, 2008. <http://www.uvasci.org/wp-content/uploads/2008/09/sci-6-report.pdf>

Digital Research Infrastructure for the Arts and Humanities (DARIAH) <http://www.dariah.eu/>

Digital Humanities Observatory (DHO) <http://www.dho.ie>

Humanities Serving Irish Society (HSIS) <http://www.hsis.ie> ICT Methods Database <http://ahds.ac.uk/ict-guides/>

Blogger Grrrrrrrrrls: Feminist Practices, New Media, and Knowledge Production

Chair: Martha Nell Smith
University of Maryland, College Park
mnsmith@umd.edu

Carolyn Guertin
University of Texas at Arlington
carolyn.guertin@gmail.com

Katie King
University of Maryland, College Park
katking@umd.edu

Marilee Lindemann
University of Maryland, College Park
mlindema@umd.edu

Ellen Moody
George Mason University
emoody@gmu.edu

In our read/write world, the Blogosphere creates a variety of semi-public and public spaces which can be used for various purposes, be they political, social, commercial, spiritual, intellectual. All of the presenters on this panel are active bloggers, engaging the blogosphere and its audiences for feminist purposes. As Marilee Lindemann states about her creative nonfiction project, *Roxie's World* (http://roxies-world.blogspot.com), her blogging creates "a place where I engage seriously with new modes of writing and critical inquiry and where I translate into a popular idiom much of what I have learned in the course of 25 years of reading, teaching, and reflecting on the politics of sex, gender, and other vectors of difference in U.S. culture." Katie King declares that in using blogs with her classes and to create intellectual community within and without her university and her usual cohort, "we use it for notices and directions to various groups: grad students and faculty in my women's studies department, and for my local LGBT book group. For a political working group on feminism and global academic restructuring we use it for document collection and web research resources, and I use blogging to give professional talks and at conferences, when I share my research on feminist transdisciplinary practices." Blogging can demonstrate and even alter web technologies like Google maps, and likewise demonstrate cognitive activities like scaling and scoping. These are all realities, metaphors and models for thinking. Blog spaces are now conceptual spaces for intellectual multi-connection and feminist transdisciplinary practice.

Yet as Ellen Moody notes, "Women are the emigrant minorities of cyberspace." Indeed, every study from 1985 through to 2003 of the World Wide Web demonstrates that it is a male construction: it was begun by men and exploits technologies many women have not been trained to use or are not comfortable using. A strongly masculinistic ethos produces and structures many regions and norms in Net life; much of cyberspace culture is still controlled and dominated by men, and reflects and encourages sexual and sheerly competitive aggression, hostility towards, and debasement of women, and disparages what's thought to be female points of view of social life. The language of the Net, the language of instructions, commands, and descriptions of computer behavior tends to be masculinist, and much of it still defines and looks at the action a user needs to take from the point of what an engineer or programmer has automatically caused to happen to the machinery from his angle. The jargon tends to be that of imagined cowboy and science fiction adventure violence: bash, kill, abort, master/slave, booting up. Much is rule- and product-outcome based. What is generalized in the command word is often not what the user imagines she is doing (type this in here), or the user's aims, but what the trained engineer thinks he has to done to make his technology respond to typed commands. Many of the instructional words refer not to the user's gesture, but to what is provided by some remote machine. The type of training someone must go through to become a computer programmer requires women to repress ways they have been encouraged to think and act, to replace picture and concrete thinking and women's metaphors with abstractions and metaphor drawn from a male point of view. Claude Levi-Strauss associated building through combining pictures with the savage mind and called it *bricolage*, a process central to the way website building has often been seen and experienced. Lest one think that these descriptions of web environments are exaggerated, all one needs to do is a quick archive search on the vituperative, frequently sexist spewings of the blogger boyz on Daily Kos (www.dailykos.com/) throughout the primary season of spring 2008. Or, one might review Carolyn Guertin's presentation at DH2007, "If I Can't Dance, I Don't Want To Be In Your Digital Revolution," in which she recounted and analyzed the "seedy underbelly of the blogosphere and cyberspace populated by Alpha Dogs, Griefers, and Trolls. **Alpha Dogs** are people who will use any form of abuse to pump themselves up or to 'win' a point; **Griefers** are people who annoy other people, and **Trolls** are people who post hate speech and inflammatory messages about a person or topic in order to bait other users into responding."

So all four panelists are acutely aware of both the great promise and hope of the blogosphere and of the fact that it is by no means an Edenic space, or at the very least there are rapacious as well as facilitating, connective machines in the new media garden. This panel's presentations will not dwell on what Guertin class the "seedy underbelly" but will propose methodologies and practices for reflecting on the important questions framing this panel, questions about how and what kind of knowledge we are producing in the wired world, how differences are negotiated and communities are created, and how agencies are located and distributed. The panel's focus will be on diversity not within digital humanities as a field but within the social and political networks being created in the blogosphere. One blog featured in the panel, *Roxie's World,* can serve as an example for the kinds of expressions explored by the panel as a whole. Queer, feminist, comedic, and utopian, *Roxie's World* is written by a prominent Americanist, feminist, queer theory scholar in the voice of a 14 year old wire haired fox terrier with a leaky heart and a laoptop. It is postmodern in its giddy appropriation and re-contextualization of images, video, words, and music, in its gleeful movement back and forth across the boundaries between the animal and the human, the fictive and the factual, the personal and the political. It is political in its determination to "talk back"—to borrow a term from bell hooks—to a culture of images and talking heads and to dissent from the authoritarianism of post-9/11 Bush/Cheneyism. It is classically, literarily "American" in its deployment of an innocent narrator as the vehicle for critique and satire, a narrator constantly surprised by the betrayals and disappointments of the adult humans in charge of the world.

Our critical inquiries are situated at the intersection of queer/feminist studies of the public sphere (Lauren Berlant, Michael Warner); an emerging discourse of blog studies (particularly two recent online collections, *Into the Blogosphere*: *Rhetoric, Community, and the Culture of Weblogs* from the University of Minnesota and *Blogging Feminism: (Web)sites of Resistance* from the Barnard Center for Research on Women); and, at least in the case of *Roxie's World,* canine cultural studies (Marge Garber, Donna Haraway, Alice Kuzniar [who gets credit for the term "Canine Cultural Studies," by the way]. Our experiments in the blogosphere have afforded us opportunities to reflect upon what blogs are, what they do—culturally, politically, literarily—and what they can teach us about practices of reading, writing, and social networking in the twenty-first century. On our blogs, we each engage in gestures of "digital self-fashioning" that we regard with a certain wariness, seeing them not necessarily as liberatory or equalizing but also as evasions, erasures, or denials of real and powerful social locations and differences. All of us in the digital humanities long ago recognized that new technologies had helped to produce forms of textuality and ways of reading and writing that are genuinely new and worth reckoning with, but, demanding participation, blogging arguably helps clarify the stakes and the possibilities of a whole new way of conceptualizing and doing humanistic work in ways that our other tools do not (or at least not so readily). As scholars, critics, teachers, and writers of various kinds, we all have a stake in getting into the game and seeing what we can do in it and with it. The tools are there. It's up to us to find innovative ways to use them, ways that will reflect our values and commitments and advance our scholarly and cultural work. It's important to do so not just to prove our relevance or to find common ground with our tech-savvy students, though those are not insignificant considerations. It's important because the transition to post-print culture is advancing by the nanosecond. As Richard Miller of Rutgers has recently pointed out, a revolution in human expression is taking place. As scholars of the culture of human expression, we simply must develop the skills necessary for interpreting and understanding emerging forms of textuality, and, as Lindemann has written, "there is a lot to be said for learning by doing, which is how I have justified my now lengthy experiment in digital creativity. In other words, and you had to know I would come to this, my old dog has taught me several new tricks, and I am grateful to her for being both a loyal companion and an extraordinary teacher."

Guertin, King, Lindemann, and Moody will examine issues of gender and other identity relations and formations, the perils and possibilities of enclaves, and the meanings of new ways and means of writing and reading to postulate ways in which blogging might be harnessed to advance knowledge production and forge new human-human and human-machine interactions. Doing so, we will probe the following in order "to rethink the intricate, the increasingly intimate, configurations of the human and the machine" and configurations of human-human interactions through machine-facilitated interactions:

- The irreducibility of lived practice, embodied and enacted;

- The value of empirical investigation in the midst of the relentless valuing of categorical debate;

- The displacement of reason from a position of supremacy to one among many ways of knowing and acting;

- The heterogeneous sociomateriality and real-time contingency of performance;

- New agencies and accountabilities effected through reconfigured relations of human and machine. (Lucy Suchman, *Human-Machine Reconfigurations* 1, xii).

Important also for our considerations are the special issue of *Frontiers* devoted to gender, race, and information technology (http://muse.jhu.edu/journals/fortiers/toc/fro26.1.html); recent books by feminist thinkers such as Isabel Zorn, J. McGrath Cohoon and William Aspray, as well as Suchman; and the recent issue of *Vectors* devoted to *Difference* (Fall 2007; http://www.vectorsjournal.org/index.php?page=6%7C1).

Working Bibliography

Beetham, Gwendolyn, and Jessica Valenti, eds. "Blogging Feminism: (Web)sites of Resistance." *The Scholar and the Feminist Online* 5.2 (2007). http://barnard.edu/sfonline/blogs/index.htm.

Berlant, Lauren. *The Female Complaint: The Unfinished Business of Sentimentality in American Culture*. Durham and London: Duke UP, 2008.

---. *The Queen of America Goes to Washington City: Essays on Sex and Citizenship*. Durham and London: Duke UP, 1997.

Curtin, Tyler. "Promiscuous Fictions." *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. Eds. Laura J. Gurak, et. al. <http://blog.lib.umn.edu/blogosphere/promiscuous_fictions.html>.

Doty, Mark. *Dog Years: A Memoir*. New York: HarperCollins, 2007.

Duggan, Lisa. *The Twilight of Equality: Neoliberalism, Cultural Politics, and the Attack on Democracy*. Boston: Beacon, 2004.

Fraser, Nancy. "Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy." *Habermas and the Public Sphere*. Ed. Craig Calhoun. Cambridge, MA: MIT P, 1992. 109-42.

Garber, Marjorie. *Dog Love*. New York: Simon and Schuster, 1996.

Guertin, Carolyn. "All the Rage: Digital Bodies and Deadly Play in the Age of the Suicide Bomber." Anthology of the Virginia Tech killings. *There's A Gunman on*

Campus: Tragedy and Terror at Virginia Tech. Ed. Ben Agger and Timothy W. Luke. 215-228. New York: Rowman & Littlefield, 2008.

Gurak, Laura J., Smiljana Antonijevic, Laurie Johnson, Clancy Ratliff, and Jessica Reyman, eds. *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. June 2004. <http://blog.lib.umn.edu/blogosphere/visual_blogs.html>.

Haraway, Donna. *The Companion Species Manifesto: Dogs, People, and Significant Otherness*. Chicago: Prickly Paradigm Press, 2003.

---. *When Species Meet*. Minneapolis and London: U of Minnesota P, 2008.

Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: U of Chicago P, 1999.

---. *Writing Machines*. Cambridge and London: MIT P, 2002.

hooks, bell. *Talking Back: Thinking Feminist, Thinking Black*. Boston: South End Press, 1989.

King, Katie. "Networked Reenactments: A Thick Description amid Authorships, Audiences and Agencies in the Nineties." *Writing Technologies* 2.1 (2008) http://www.ntu.ac.uk/writing_technologies/Current_journal/King/index.html.

Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT P, 2008.

Kuzniar, Alice. *Melancholia's Dog: Reflections on Our Animal Kinship*. Chicago and London: U of Chicago P, 2006.

Latour, Bruno. *We Have Never Been Modern*. Cambridge, MA: Harvard UP, 1993.

Lovink, Geert. "Blogging, the Nihilist Impulse." *Eurozine*. Jan. 2, 2007. http://eurozine.com/articles/2007-01-02-lovink-en.html.

McGill, Meredith. "Lurking in the Blogosphere of the 1840s: Hotlinks, Sockpuppets, and the History of Reading." *Common-Place* 7.2 (2007). http://common-place.org/vol-07/no-02/reading/.

Margolis, Jane and Allan Fisher, *Unlocking the Clubhouse: Women in Computing*. Cambridge, Massachusetts and London, England: The MIT Press, 2003.

Osell, Tedra. "Where are the Women?: Pseudonymity and the Public Sphere, Then and Now." "Blogging Feminism: (Web)sites of Resistance." *The Scholar and the Feminist Online* 5.2 (2007). Eds. Gwendolyn Beetham and Jessica Valenti. http://www.barnard.columbia.edu/sfonline/blogs/osell_01.htm.

Perlmutter, David D. *Blogwars: The New Political Battleground*. New York: Oxford UP, 2008.

Povinelli, Elizabeth A. *The Empire of Love: Toward a Theory of Intimacy, Genealogy, and Carnality*. Durham and London: Duke UP, 2006.

Roberts-Miller, Trish. "Parody Blogging and the Call of the Real." *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. Eds. Laura J. Gurak, et. al. http://blog.lib.umn.edu/blogosphere/parody_blogging.html.

Serfaty, Viviane. *The Mirror and the Veil: An Overview of American Online Diaries and Blogs*. Amsterdam and New York: Rodopi, 2004.

Smith, Martha Nell. "Software of the Highest Order: Revisiting Editing as Interpretation." *Textual Cultures* 2.1 (Spring 2007): 1-15.

Suchman, Lucy A. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge: Cambridge UP, 2007.

Sunstein, Cass R. *Infotopia: How Many Minds Produce Knowledge*. New York: Oxford UP, 2008.

---. *Republic.com 2.0*. Princeton: Princeton UP, 2007.

Warner, Michael. *Publics and Counterpublics*. New York: Zone Books, 2002.

Supporting the Digital Humanties: putting the jigsaw together

Chair: Martin Wynne
University of Oxford
martin.wynne@oucs.ox.ac.uk

Steven Krauwer
Utrecht University (CLARIN)
steven.krauwer@let.uu.nl

Sheila Anderson
King's College, London (DARIAH)
sheila.anderson@kcl.ac.uk

Chad Kainz
University of Chicago (Project Bamboo)
cjkainz@uchicago.edu

Neil Fraistat
University of Maryland (CenterNet)
fraistat@mac.com

This panel session will present various important current international initiatives which aim to support and take forward research in the digital humanities. In particular, it will identify the various problems that these different initiatives aim to address, compare the different approaches, and seek to draw out the connections, synergies and potential areas of overlap and competition in trying to build a coordinated environment for the next generation of advanced research.

There is a growing awareness and agreement that the principal problem facing the digital humanities is the fragmented nature of the research environment. For example, numerous tools exist, but are not interoperable with each other, or with the many datasets that continue to be prepared according to varying standards and made available according to various licensing agreements and access arrangements. This fragmentation and lack of interoperability has dogged the digital humanities for many years, and is limiting the impact of digital resources and tools on mainstream research in the humanities.

Numerous initiatives in the past have attempted to address some of the individual aspects of the problems: to provide better tools for analysis, to create resources, to offer resource discovery services, to set up repositories, to set up networks of expert centres, to define and promote standards and good practice, to develop tools to support collaboration, to offer advisory and support services, and to harness the new paradigms of 'cyberinfrastructure' and 'e-Science' for the benefit of the Humanities. However, it has become clear that a number of more or less unconnected initiatives in these individual areas cannot address the overall problem of fragmentation and lack of coordination.

As the recognition spreads that the number and heterogeneity of the varying services, tools, and standards that are holding us back, a new round of initiatives are aimed overcoming the problem of fragmentation, by promoting concerted, coordinated, international efforts to build a sustainable infrastructure to support and sustain effect research in the digital humanities. At the national level in several countries, infrastructure is now on the agenda, either specifically for the Humanities, or in cross-disciplinary initiatives which now include the Humanities. In Europe, the European Strategy Forum on Research Infrastructures now includes the humanities in its roadmap, and the CLARIN and DARIAH projects are underway to prepare to construct research infrastructures, and other regional structures are starting to emerge. At the international level, CenterNet is now established as a network and forum for expert centres in the digital humanities, and Project Bamboo is proposing to build a new generation of tools to support research, responding to the real and specific user requirements in the arts and humanities.

There is now an urgent priority for these initiatives to communicate, to share ideas and experiences, and to investigate the possibilities of coordinating their activities.

This panel will address the ways in which some of the most important current international initiatives are attempting to address these problems. Each speaker will be asked to address the following questions:

- What specific problems have you identified, and how is your initiative seeking to address them?

- What services, if any, do you ultimately aim to provide?

- How might you link with other related initiatives?

- What are the further elements of the jigsaw puzzle which are needed to create a coordinated and more complete research infrastructure?

Papers

Digital and Virtual Architecture: a review of two projects

Nicoletta Adamo-Villani
Purdue University
nadamovi@purdue.edu

1. Introduction

During the past four decades digital technologies have had a major impact on architecture and have completely revolutionized the way architecture is designed and visualized. The ability to digitally manipulate architectural components and to consider all possible configurations in advance has provided new, alternate means to design architecture (Bermudez & Klinger, 2003). If we observe some examples of "digitally designed architecture", such as the Hadid nuragic and contemporary art museum in Cagliari (Zaha Hadid, 2007) and Frank O. Gehry's Experience Music Project (O. Gehry, 2000) we note that these unique architectural designs could be created only in a computer-mediated environment.

Digital representation has proven to be an excellent tool for the communication of design ideas (Mohan, 2003). Instead of symbolic representations, static renderings and time consuming physical models, technologies such as photorealistic 3D animation have made it possible to create convincing renderings and walkthroughs that can be easily understood by non-specialists (Uddin, 1999).

In addition, a completely new form of computer-based architecture has recently emerged: Virtual Architecture. Maher defines Virtual Architecture as an '..interactive, networked spatial environment designed using the metaphor of physical architecture, from which it inherits many visual and spatial characteristics' (Maher et al., 2000). The main purpose of Virtual Architecture is to provide an electronic location for people to socialise, work, and learn in the same way physical architecture does. Because Virtual Architecture uses the metaphor of cities, buildings and rooms, it can be designed by architects and then constructed by computer graphics programmers (Maher et al., 1999).

In this paper we describe and discuss two architectural projects that were developed entirely in a computer-based environment. The first is a 3D animation-based project whose goal was to design and visualize a visionary new city. The second project is an example of Virtual Architecture and its objective was to design and develop an online virtual world.

2. "Samarkand on the Euxine": a Digital Architecture project

The objective of the project was to design and visualize a new city in the Republic of Turkey near the present site of Istanbul. The city would serve as a satellite city that would provide aid in the event of a catastrophic earthquake. Devastating earthquakes have visited Istanbul in the past at intervals varying from 300 to 100 years and the next one is due any moment. The Metropolitan Municipality of Istanbul is planning for response to the dreaded event in many ways. One of these is the development of an entirely new satellite city.

The vision of the city was created at the Envision Center at Purdue University by a team of graduate and undergraduate students led by the author and by a faculty in Purdue Civil Engineering. The team chose to use photorealistic 3D animation technology because of its ability to represent design in an intuitive, concrete way. The main advantage of 3D photorealistic animation over CAD visualizations is 'representational fidelity'. Representational fidelity refers to the degree of realism of the rendered 3-D objects and the degree of realism provided by temporal changes to these objects. 3D animation-based visualizations include true-to-life replicas of buildings, displayed with accurate perspective, occlusion, photographic-quality materials, lights, and motion. When making a presentation supplemented with a photorealistic 3D visualization, architects and designers can convey their visions clearly to those who need to understand them in order to make decisions. In this case, it is hoped that the powerful 3D imagery produced will convince the decision-makers at the Metropolitan Municipality of Istanbul to act with the needed alacrity.

The project presented many challenges because of its complexity, the scale of the area to be visualized (40,000 acres), and the aggressive production timeline. The overall city plan was developed with the goal of creating a green city with an advanced cyber infrastructure dedicated to communications, security, and recycling. The plan includes a business district, research and government centers, modern museums, concert halls, theaters, hospitals, retail centers, exhibition halls, buildings for social functions, a sports center, and a hotel district. Figure 1 shows a partial 3D map of the city.

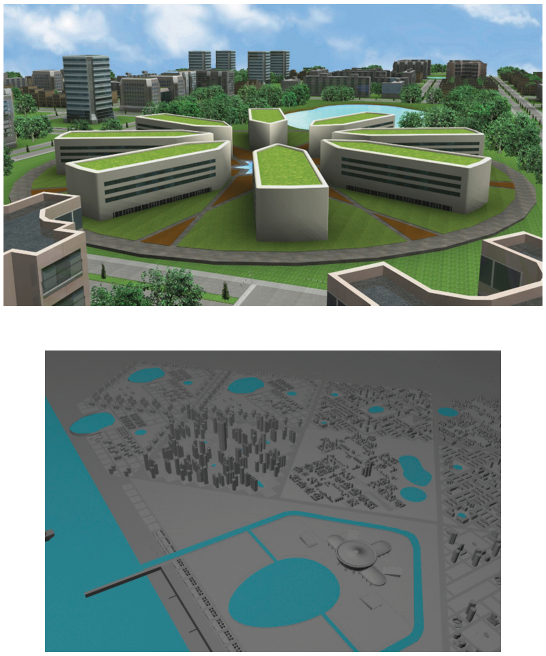


Figure 1. 3D map; “Selcuk Center”

The design started with a low-detail city plan in the form of a .dwg AUTOCAD file. The majority of the design decisions were made directly in the 3D software as sections of the city were being visualized. The main design goal was to create a modern city that takes advantage of state-of-the-art building techniques and, at the same time, includes architectural elements reminiscent of the Byzantine and Ottoman Empires. This design theme is clearly evident in Figure 1 which shows a cluster of earthquake-resistant buildings arranged in the shape of the “Selcuk star”, a classical Ottoman symbol.

All buildings were designed, modeled and textured in MAYA 8.5 software. Global illumination and radiosity were used to give an accurate depiction of real life lights and environments; photo realistically rendered human figures, vehicles and landscape elements were added in order to accentuate the realistic look of the visualization and provide a sense of scale. The animation can be viewed at: http://www.tacc.utexas.edu/~jwozniak/TG08/13/13.html“

3. The 21st Century World”: a Virtual Architecture project

The “21st Century World” is a collaborative project between Purdue University and Educate for tomorrow (EforT, Hawaii). Its objective is the development of a 3D online virtual city designed to allow students and the general public to interactively explore nanotechnology enhanced products of the 21st Century. Users can travel through the city, enter buildings, manipulate objects, and interact with 3D avatars to learn about nanotechnology. The virtual city includes buildings with nanoenhanced materials and self cleaning windows that absorb energy from sun light. It features cleaner and safer mass transportation, alternate fuel stations, cars with self-repairing body-paint scratches, and buildings that grow crops indoors.

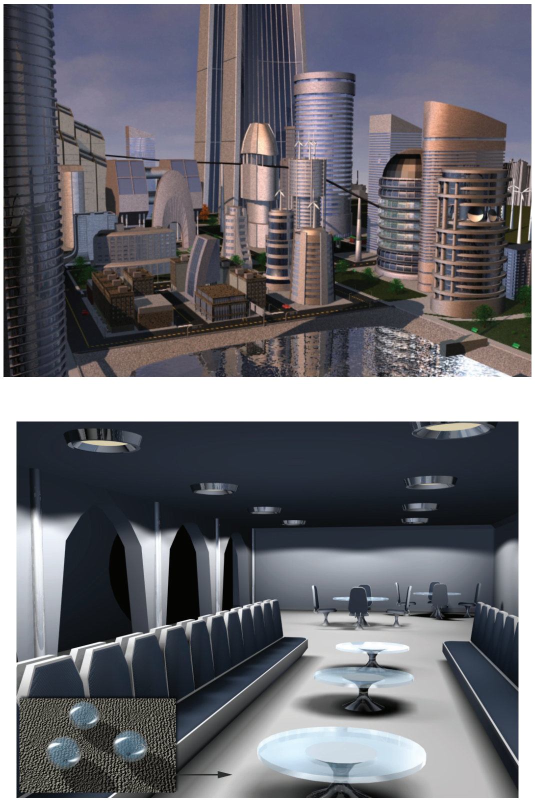


Figure 2. Renderings showing part of the 21st Century World and interior of train station with detail of water and dirt repellent floor

The futuristic city was designed entirely in 3D STUDIO MAX 9.0 software by a team of architects and nano-science experts (figure 2 shows two renderings of the city). The models were exported from 3D STUDIO MAX as VRML files and the interactive application is currently being programmed. Images and fly-trough animation can be accessed at: http://www2.tech.purdue.edu/cgt/i3/nanofactor/web%20site/index.htm

4. Discussion

The project described in section 2 demonstrates how digital technologies such as photorealistic 3D modeling and animation can benefit architectural visualization and design. The primary advantage of 3D photorealistic modeling lies in the ability to visualize design in more concrete terms. The usefulness of maps and architectural drawings can be limited, particularly to the non-specialist, whereas a photorealistic virtual model can enable decision makers and the general public to see the aspects of a project from every possible angle. 'As a result, it can increase the level of feedback and constructive responses before taking any financial risks.' (Wilson, 1998). In our case, clear communication and risk mitigation were the most compelling drivers of incorporating 3D photorealistic modeling/animation into the project. Considering the cost, scale and scope of projects like the one described in the paper, producing a photorealistic virtual model is a good investment because it can speed up decision making processes, streamline approvals, and reduce costly change orders (Bouchlaghem, 2005).

Moreover, the ability to interactively experiment with a variety of shapes and forms and see the results being visualized in real-time time are tremendous benefits of 3D modeling and animation technology. Samarkand on the Euxine has received very positive feedback from the officials at the Municipality of Istanbul and has therefore demonstrated that effective architectural designs can be created entirely in a computer-mediated environment. The author agrees with (Bermudez & Klinger, 2003) who argue that digital thinking is indeed architectural thinking.

The project described in section 3 is an example of virtual place that draws on knowledge of architectural design. Maher argues that most virtual environments are created by programmers rather than designed by architects and, as a result, we are in the "era of vernacular virtual architecture" (Maher et al., 2000). In the development of the 21st Century World, provision of functionality and geometric description of the space were considered equally important factors. The objective was not only to create a highly interactive and functional environment, but also to design a city based on the principles of good urban design. In particular we considered the principle of 'Legibility and Wayfinding' to help the users in orientation and navigation tasks; the principle of 'Character and Meaning' to help the participants recognize and value the differences between one area and another; and the principle of 'Order and Incident' (e.g., balancing consistency and variety) to provide the users with an appealing environment that promotes curiosity and motivates them to continue to explore (Barnett, 1982) (Larice & MacDonald, 2007). 27 users have evaluated the 21st Century World so far. Their feedback on usability and appeal has been extremely positive, thus confirming the importance of implementing architectural and urban design principles in the design of virtual places.

References

Barnett, J. (1982). *An Introduction to Urban Design*. Harper & Row, New York .

Bermudez, J. & Klinger, K. (editors) (2003). Digital Technology and Architecture - White Paper. *ACADIA 2003*. http://www.acadia.org/ACADIA_whitepaper.pdf.

Bouchlaghem, D., Shang, H., Whyte, J., Ganah, A. (2005). Visualisation in architecture, engineering and construction (AEC). *Automation in Construction*, 14, 287– 295.

Frank.O.Gehry's Experience Music Project. http://www.arcspace.com/architechts/gehry/emp_n/

Larice, M. & MacDonald, E. (editors) (2007). *The Urban Design Reader*. Routledge, New York London.

Maher, M.L., Simoff, S., Gu, N., Lau, K.H. (2000). Designing Virtual Architecture. *Proceedings of CAADRIA 2000*, Singapore.

Maher, M.L., Gu, N., Li, F. (1999). Visualisation and Object Design in Virtual Architecture *Proceedings of CAADRIA 1999*, Sydney, Australia.

Nethra Mettuchetty Ram Mohan (2003). Emerging Technologies in Architectural Visualization – Implementation Strategies for Practice. MS Thesis - School of Architecture Mississippi State. http://sun.library.msstate.edu/ETD-db/theses/available/etd-04072003-164447/unrestricted/nethra_thesis.pdf.

Uddin, M. S. (1999). *Digital Architecture*. McGraw- Hill

Wilson, J.D. (1998). ModelCity Philadelphia Elevates GIS to the Next Level. *Professional Surveyor*, 18(2).

Zaha Hadid nuragic and contemporary art museum in Cagliari, Italy. http://www.arcspace.com/architects/hadid/cagliari/cagliari.html.

Patterns in Style Evolution of Poets

Vadim Sergeevich Andreev
Smolensk State University
smol.an@mail.ru

Quantitative studies of the evolution of style ('stylochronometry') have revealed systematic changes in the features (syntactic, morphological, lexical, etc.) of texts of the same author over time (Goldfield, Hoover 2008; Juola 2008; Stamou 2008). These findings stimulate further research aimed at detection of the individual and universal markers of style evolution of different authors and, on the other hand, integral features, ensuring style contiguity, at establishing factors which influence style development in different genres and other issues.

The aim of this presentation is to address some of the above-mentioned problems on the material of verse: to find out the pattern of the evolution of the style of American romantic poet H.W. Longfellow and to compare it to the development of the style of another famous romantic poet E.A.Poe. Poe's style evolution was analyzed in our previous research (Andreev 2008).

Choosing the material for the study we tried to meet the requirements for the research in stylometry (Rudman 2003). Texts which were selected for our analysis were of the same genre and comparable in size – iambic lyrics not exceeding 65 lines. We analyzed the poems which the authors themselves included into their collections of works due to which they reflect the authors' subconscious stylistic preferences. The possibility of serious editorial interventions or any other unauthorized alterations is reduced to the minimum.

In stylochronometric research the feature set, used by different scholars, varies considerably (Juola 2008; Stamou 2008). In our study we use 29 characteristics which were selected as a result of a number of experiments. They include 10 morphological and 19 syntactic features. Morphological characteristics reflect the frequency of different morphological classes of words (noun, verb, adjective, adverb and pronoun) in the first and final strong (predominantly stressed) syllabic positions – ictuses. Most syntactic characteristics are based on the traditional notions of the members of the sentence (subject, predicate, object, adverbial modifier) in the first and the final strong positions in the lines. Other syntactic parameters are the number of clauses in (a) complex and (b) compound sentences, inversions (a) complete (inversion of the subject or predicate) and (b) partial (inversion of the other members of the sentence), lines divided by syntactic pauses, enjambements (absence of syntactical pause between the lines) and lines ending in exclamation or question marks.

The creative life of Longfellow in our study is divided into three periods: early, middle and final. In order to determine the borderlines between the periods biographical data was used. The first period embraces the years of his youth (1819 to 1825) till Longfellow started his working career. During this period Longfellow wrote lyrics which he later included into his two collections *Earlier Poems*, *Juvenile Poems*. The beginning of the last period is marked by the Civil War, which according to the majority of critics became a turning point in the development of American romanticism. This period is represented by his collections *Flower-de-Luce* and *Ultima Thule*. The remaining part of Longfellow's creative career constitutes the second period (collections *Voices of the Night*, *Ballads and Other Poems*, *Poems on Slavery*, *The Belfry of Bruges and Other Poems* and *The Seaside and the Fireside*).

To establish parameters differentiating the texts of the above-mentioned three periods and to estimate their discriminant force multivariate discriminant analysis was used (Warner 2008: 650-701). The results of this analysis show that there are substantial and systematic alterations in the style of Longfellow over time. Out of 29 characteristics 15 were found to possess discriminant force, the most relevant being the number of subordinate clauses, enjambements, lines divided by syntactic pauses, inversions and the number of nouns, verbs and adjectives in the initial strong position in the line. Post hoc success classification rate is nearly 97%.

Since the second period is longer than the other two we tested whether this can influence the results (Rybicki 2008). During this experiment instead of all texts, belonging to Period 2, a 50% random sample of the texts was investigated, while the remaining poems were used as a control group for a cross-validation test.

This additional experiment gave very similar results revealing the same pattern in the poet's style evolution. Success classification rate now was 90% and the cross-validation test yielded 64,70% of correct period prediction, which is two times higher than chance level. Such results can be considered as acceptable (Juola 2008, 288-291).

Earlier lyrics by Longfellow (Period 1) are characterized by frequent use of subordinate clauses, adjectives in the initial part of a line, inversions and enjambements. The following example is a typical stanza of young Longfellow's poems:

When the bright sunset fills,
The silver woods with light, the green slope throws
Its shadows in the hollows of the hills,
And wide the upland glows.
(An April Day)

The second period is characterized by fewer subordinate clauses, enjambements and inversions. In the initial part of the line verbs become more frequent. In the final position of the line the growth of the number of words in the syntactic function of attribute is observed. Some of these changes can be illustrated by the following example:

A new Prometheus, chained upon the rock,
Still grasping in his hand the fire of Jove,
It does not hear the cry, nor heed the shock,
But hails the mariner with words of love.
(The Lighthouse)

The third period is marked by further decrease in the use of subordinate clauses and the growth of the number of enjambements, partial inversions and lines divided with syntactic pauses. The number of verbs in the first stressed position decreases and the number of nouns increases. Many features of the poet's style resemble those of his earliest poems:

The wind blows, and uplifts thy drooping banner,
And round thee throng and run
The rushes, the green yeomen of thy manor,
The outlaws of the sun.
(Flower-de-Luce)

As we see, morphological characteristics of the end of the line are the most stable. On the contrary, there is a clearly marked change in the author's use of morphological classes of words in the initial strong position: adjectives in the first period – verbs in the second period – nouns in the third one. Syntactic features, which make verse more complex and expressive, such as enjambement, syntactic pauses in the lines and inversions, are frequent during the first period, then there is a drop in their use during the second period and at the end of his life the author resumes their frequent use.

The comparison of style evolution of such different poets as Longfellow and Poe reveals not only differences but, quite unexpectedly, certain similar tendencies, too.

Major differences can be summed up as follows:

- in Poe's lyrics morphological characteristics of both initial and final parts of the line change over time – in Longfellow's texts the end of the line is stable;

- the increase of verse expressiveness at the level of syntax for Poe is a constant tendency and in case of Longfellow the frequency of expressive syntactic characteristics decreases at the second stage and rises again in the third period.

Major similarities:
Despite the fact that the poets belonged to different schools of American romanticism there are similar patterns in their style development.

- The number of subordinate clauses gradually reduces over time.

- The evolution of style of both poets was not rectilinear: certain deviation from the initial syntactic pattern of verse line is observed during the middle period of their creative activities and then at the final stage both authors resume, to a large extent, the use of the features, typical of their early lyrics.

References

Andreev V. (2008) Variation of Style: Diachronic Aspect. *Digital Humanities 2008. Conference Abstracts*: 42-43.

Goldfield J., Hoover D.L. (2008) Homebodies and Gad-Abouts: A Chronological Stylistic Study of 19th Century French and English Novelists. *Digital Humanities 2008. Conference Abstracts*: 117-120.

Juola P. (2006) Authorship Attribution. *Foundations and Trends in Information Retrieval*. Vol. 1, No.3: 233-334.

Rudman J. (2003) Cherry Picking in Nontraditional Authorship Attribution Studies. *Chance*, Vol.16, No. 2: 26-32.

Rybicki J. (2008) Does Size Matter? A Re-examination of a Time-proven Method. *Digital Humanities 2008. Conference Abstracts*: 184.

Stamou C. (2008) Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating. *Literary & Linguistic Computing*, 23: 181–199.

Warner R.M. (2008) *Applied Statistics*. Los Angeles – London: Sage Publications, 2008.

An Approach to Treating Videos as Academic Documents

Stewart Arneil
University of Victoria
sarneil@uvic.ca

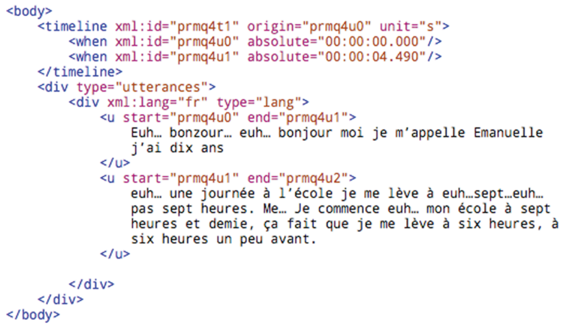
Greg Newton
University of Victoria
gregster@uvic.ca

As the scope of what constitutes a text expands to include time-oriented media such as video, it is becoming increasingly popular to incorporate multimedia in to the contexts of research and teaching. As audiences become more discriminating so too have designers and developers begun to push the limits of technology to accommodate these new appetites for more media, more features, more collaboration- more everything. Applications to collect and disseminate media to an engaged audience are becoming so sophisticated that we have entered in to a realm where our technical reach sometimes outstrips the users' capacity to imagine its use; a situation occasionally referred to as an answer without a question. But we did have a question: how best to take several hundred videotapes of high-profile guest lecturers collected over the course of 20 odd years and turn them into an engaging, easy to use webbased application that provided users with more than just a talking head?

Our office provides specialized software development service to, and collaborates in research with, faculty members in at least a dozen disciplines. Our preference is to abstract from the immediate task and look for ways to make our work extensible or transferable to other projects. What we were looking for was a general purpose tool to apply to videos to help with what John Unsworth calls "scholarly primitives": Discovering, Annotating, Comparing, Referring, Sampling, Illustrating, Representing. (Unsworth, 2000)

Our initial objective was to provide simultaneous transcription and simple valueadded features; ancillary information in the form of links to other sites and images germain to the current utterance. We considered existing technologies for marking up and presenting videos such as those in use at MIT's OpenCourseWare <http://ocw.mit.edu/> and TED <http://www.ted.com>) but they did not provide anything more sophisticated than metadata and full text search, let alone provide for multiple channels of time-related support material. On the other hand, our users were not interested in highly detailed performance attributes, such as those described by Saltz (Saltz 2004). We looked at SMIL, but quickly discarded it in light of the immaturity of existing playback systems. It did, however, provide us with a model. XML is a heavily used technology in our shop and we were able to produce a TEI schema for encoding transcripts based on conventional semantics of utterances: we specifically used the Transcriptions of Speech module to encode all information.

Although there was a paucity of existing software that would provide us with an inclusive playback mechanism it was not beyond the project's scope to produce our own. Indeed, the more we looked at our needs, the more reasonable it seemed. Our specification ended up being rather short: XML would provide the natural structure that such texts demand; multi-modal data streams would remain separate both in terms of storage and delivery, allowing us to abstract code such that we could remove any dependence upon a single media player; users should also be able to bookmark, and therefore cite, specific points in the video.



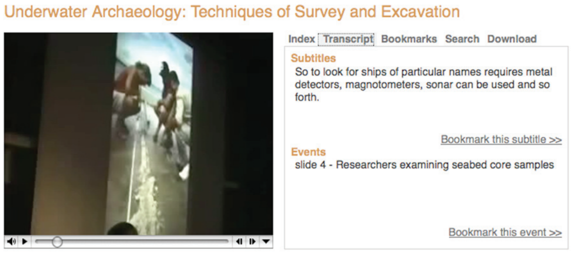
```
<body>
    <timeline xml:id="prmq4t1" origin="prmq4u0" unit="s">
        <when xml:id="prmq4u0" absolute="00:00:00.000"/>
        <when xml:id="prmq4u1" absolute="00:00:04.490"/>
    </timeline>
    <div type="utterances">
        <div xml:lang="fr" type="lang">
            <u start="prmq4u0" end="prmq4u1">
                Euh… bonzour… euh… bonjour moi je m'appelle Emanuelle
                j'ai dix ans
            </u>
            <u start="prmq4u1" end="prmq4u2">
                euh… une journée à l'école je me lève à euh…sept…euh…
                pas sept heures. Me… Je commence euh… mon école à sept
                heures et demie, ça fait que je me lève à six heures, à
                six heures un peu avant.
            </u>

        </div>
    </div>
</body>
```

Figure 1. A representative snippet of XML from the project in an editor

Implementation, then, would focus on treating the document as an academic "paper", with a feature set that anticipates the expectations and requirements of a scholarly user. Each timeline (transcript, events, commentary) consists of a list of when elements; each when element identifies timestamps in the video and relates them to xml elements in the file. The XML files are stored in an XML data base (eXist), which allows for highly sophisticated xqueries if necessary. Identifying the elements in the video stream and marking up the support documents are done manually with commercial video playback and XML editors (see figure 1). Our approach is complementary in many ways to that of the AXE project (Reside, 2007).

A proof-of-concept was constructed using PHP and relying on the QuickTime player, due to its rich javascript API. As QuickTime announces its play head position the page determines which utterance in each timeline is current and displays a quickly digestible block of text to the viewer for each timeline (see figure 2).


Underwater Archaeology: Techniques of Survey and Excavation
Index
Transcript
Bookmarks
Search
Download
Bookmark this subtitle >>
Bookmark this event >>

Subtitles

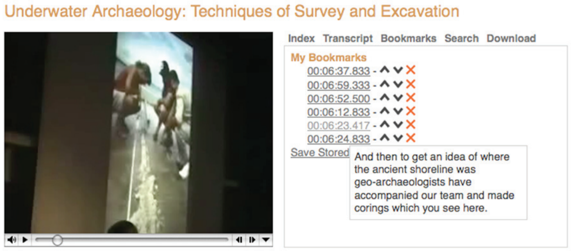
So to look for ships of particular names requires metal detectors, magnotometers, sonar can be used and so forth.

Events

slide 4 - Researchers examining seabed core samples

Figure 2. The media player and two concurrently updated channels of information

Any given utterance can be bookmarked and stored for later retrieval, providing a pinpoint-accurate citation which addresses the video itself; a "video quote" if you will (see figure 2). In addition, when the user hovers over the bookmark, the text of the utterance appears. The entire corpus or a specific video are fully searchable, with results being displayed as direct links in to the video. The same interface conventions are used for the search feature. Alternative views of the transcript are also available, including viewing the entire text on-screen, or choosing a XHTML or XML (P5 TEI) download.



My Bookmarks

00:06:37.833 - ^ ⌄ ✕
00:06:59.333 - ^ ⌄ ✕
00:06:52.500 - ^ ⌄ ✕
00:06:12.833 - ^ ⌄ ✕
00:06:23.417 - ^ ⌄ ✕
00:06:24.833 - ^ ⌄ ✕

And then to get an idea of where the ancient shoreline was geo-archaeologists have accompanied our team and made corings which you see here.

Fig. 3 The media player and bookmarks

This small proof-of-concept was modified slightly for a French instructor. She wished to assemble a number of short videos of French-speaking people from around the world with two goals: to improve students' francophone cultural literacy, and to have a corpus amenable to various kinds of linguistic research. (Caws, 2007) We continue to add videos to this collection as we obtain them from around the world.

Our original task of "rescuing" the hundreds of videotaped lectures has been scaled up to become a real project, encouraging us to re-imagine the possibilities and address the shortcomings of our proof-of-concept application. We see version 2 including a more refined feature set, with the codebase moving from PHP to Cocoon in order to improve the portability and modularity of the system. Refinements will include an online system for writing transcriptions and reducing our dependency on media players by utilizing new features in HTML5. This functionality can also be used to provide an annotative channel that is accessible to all users. Storage and "playback" of annotative snippets can provide a rich layer of added value without incurring large investments of development time because it recycles the immensely useful transcription code; this wiki-like feature has obvious value in both teaching and research contexts.

Working backward from the concept of the "video quote" we began to imagine a context in which an entire corpus of video texts is peer-reviewed and published online. Not original, unfortunately (http://www.jove.com/, http://www.vjortho.com/), but the recent emergence of such journals indicates that we may all be re-thinking what constitutes an academic journal. We do not envisage a simple transplant of the format in use by the above journals. Rather, we will have to extend it to include rich layers of annotative, transcriptive and ancillary information which can provide a discoverable and searchable corpus of texts which can be referenced on a granular level, thus providing the opportunity to sample, compare and "mash up" a corpus of video-based documents to meet either research or instructional goals.

References

Arneil, S (2007) *Francotoile, Your Gateway to Francophone Culture* http://francotoile.uvic.ca/search.php (accessed 1 Nov 2008).

Caws, C. and Arneil, S. (2007). FrancoToile: a digital video library to develop cultural literacy in French. In C. Montgomerie & J. Seale (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007* (pp. 1989-1994). Chesapeake, VA: AACE.

Elkink, M (2006) *The Lansdowne Lectures Online*, http://lettuce.tapor.uvic.ca/~taprlans/ (accessed 1 Nov 2008).

Reside, D. (2007). *The AXE Tool Suite: Tagging Across Time and Space* http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=145 (accessed 1 Nov 2008).

Saltz D Z (2004). Performing Arts. In Susan Schreibman, Ray Siemens, John Unsworth (eds), *A Companion to Digital Humanities*. Oxford: Blackwell, 2004. http://www.digitalhumanities.org/companion/view?

docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-11&toc.depth=1&toc.id=ss1-2-11&brand=default (accessed 1 Nov 2008).

Unsworth, J. (2000). *Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?*. http://jefferson.village.virginia.edu/~jmu2m/Kings.5-00/primitives.html (accessed 1 Nov 2008).

WikiPhiloSofia: Extraction and Visualization of Facts, Relations, and Networks Concerning Philosophers Using Wikipedia

Sofia J. Athenikos
Drexel University
sofia.j.athenikos@acm.org

Xia Lin
Drexel University
xlin@cis.drexel.edu

Due to its unique structural features and rich user-generated content, Wikipedia is being increasingly recognized as a useful knowledge resource that can be exploited for various applications. Nevertheless, the mode of information search and retrieval on Wikipedia remains that of conventional keyword-based search and retrieval of a list of articles ranked in terms of keyword matching. The objective of the ongoing project reported in this paper is to create a Web-based knowledge portal using the data extracted from Wikipedia, for the sake of enabling semantics-based search and exploration. The methodology is currently applied to the philosophy domain. Hence the project name: WikiPhiloSofia. In this paper we present extraction and visualization of the facts, relations, and networks involving 300 major philosophers as obtained from Wikipedia and we partially compare the results against those obtained by using Thomson Reuters' Arts & Humanities Citation Index data. Insofar as the work aims at enabling semantics-based search and exploration by exploiting the user-generated content, it embodies the movement toward Web 3.0, i.e., the convergence of the Social Web (Web 2.0) and the Semantic Web. Insofar as it emphasizes information visualization for the sake of enhancing the user's information seeking experience at the aesthetic level as well as at the cognitive level, however, the work also embodies the trend toward information aesthetics. As such, the WikiPhiloSofia project serves as a venue of the convergence of arts, humanities, and computer/information science/technology, contributing to a paradigm shift toward the next generation of online search, retrieval, and delivery of digital information in the humanities.

Introduction

Philosophy (and here we mean Western philosophy), dating back at least to approximately 600 BC, is one of the oldest of all academic disciplines and is, in particular, one of the core disciplines in the humanities. Partly due to its long history, and partly due to the nature of the discipline itself, the domain of philosophy presents a rich semantic network involving an extended genealogy of philosophers and related philosophical concepts, ideas, and doctrines, which can be explored and examined from diverse perspectives.

Wikipedia (http://www.wikipedia.org) is an open-access, collaborative Web encyclopedia project, initiated by Jimmy Wales and Larry Sanger in 2001. Wikipedia has since grown rapidly to become one of the most sought-after resources on the Web. Due to its collaborative way of construction and due to its impressive size and growth rate, Wikipedia is considered a foremost example of Web 2.0 applications.

The objective of the ongoing project reported here, entitled WikiPhiloSofia (formerly known as The WikiPhil Portal), is to extract, analyze, and visualize meaningful and interesting facts, relations, and connections among philosophers and philosophical concepts via the automatic processing of the structural features and semantic content of Wikipedia. By doing so, we aim at creating a useful and user-friendly portal for students of philosophy as well as the general public, thereby contributing to the cause of digital humanities. The project is still in its early stage. However, the current paper extends our previous work on the project (Athenikos and Lin, 2008), especially by presenting new results and findings. Specifically, the paper illustrates extracting and visualizing the data concerning facts, relations, and networks involving 300 philosophers, as extracted from Wikipedia, for semantics-based search and exploration, and it presents a partial comparison of the Wikipedia data against those from Thomson Reuters' Arts & Humanities Citation Index (http://thomsonreuters.com/products_services/scientific/Arts_Humanities_Citation_Index).

Wikipedia Data Extraction

A prototype system for the project was implemented in Java using the Java servlet technology. First we describe the materials, methods, and results involving data extraction.

Materials

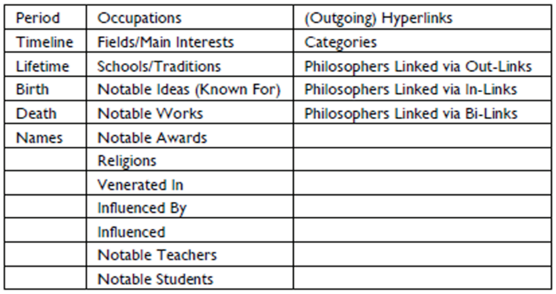
We only used the English version of Wikipedia. The initial results we had obtained in the project (Athenikos and Lin 2008) were based on the data extracted from Wikipedia pages downloaded on 29 May 2008. The results we report in this paper are based on the Wikipedia pages downloaded more recently on 23 December 2008.

Methods

We obtained a chronological list of 300 philosophers (including influential theologians, writers, scientists, etc.) from Wikipedia's "Timeline of Western Philosophers" page. We extracted information on the hyperlink connections and academic/biographical facts concerning the philosophers (as presented in infoboxes and wikitables) from their individual Wikipedia article pages, and stored the data in the form of semantic triples (Subject–Predicate–Object) in a MySQL database. We retrieved the data needed for visualization by querying the database, and stored the results as XML files marked up with GraphML and TreeML.

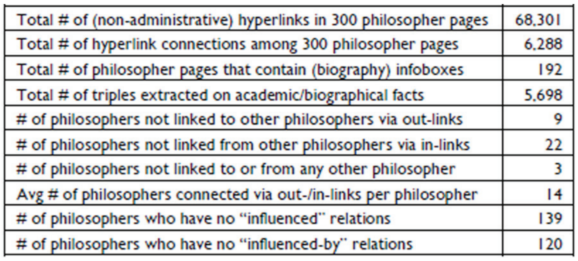
Results

Table 1 shows the types of information extracted. Table 2 summarizes the basic statistics concerning the dataset.



Period	Occupations	(Outgoing) Hyperlinks
Timeline	Fields/Main Interests	Categories
Lifetime	Schools/Traditions	Philosophers Linked via Out-Links
Birth	Notable Ideas (Known For)	Philosophers Linked via In-Links
Death	Notable Works	Philosophers Linked via Bi-Links
Names	Notable Awards	
	Religions	
	Venerated In	
	Influenced By	
	Influenced	
	Notable Teachers	
	Notable Students	

Table 1. Types of information extracted.



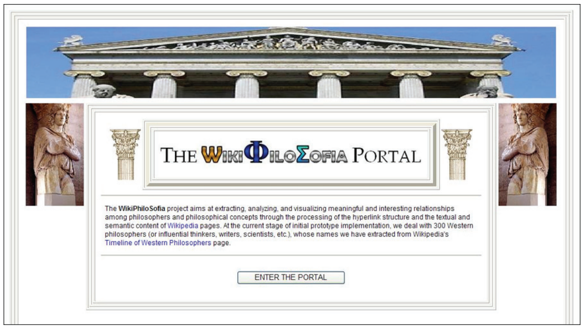
Total # of (non-administrative) hyperlinks in 300 philosopher pages	68,301
Total # of hyperlink connections among 300 philosopher pages	6,288
Total # of philosopher pages that contain (biography) infoboxes	192
Total # of triples extracted on academic/biographical facts	5,698
# of philosophers not linked to other philosophers via out-links	9
# of philosophers not linked from other philosophers via in-links	22
# of philosophers not linked to or from any other philosopher	3
Avg # of philosophers connected via out-/in-links per philosopher	14
# of philosophers who have no "influenced" relations	139
# of philosophers who have no "influenced-by" relations	120

Table 2. Basic statistics on the Wikipedia dataset.

As shown in Table 2, while there exists a high number of hyperlink connections among the 300 philosopher pages, there are a few pages that do not contain any out-links to the other philosopher pages and/or do not receive any in-links. Also, only 192 philosopher pages contain infoboxes that summarize academic/biographical facts, which is (at least in part) why there are some philosophers that are shown to have no influenced/influenced-by relations.

Semantic Search Interface

We created a Web portal interface via which the user can issue queries on the facts, relations, and networks involving 300 philosophers and explore the results displayed using diverse modalities of interactive information visualization as will be illustrated in the next section. Figure 1 shows the homepage of the WikiPhiloSofia portal (http://research.cis.drexel.edu:8080/sofia/WPS/).



The **WikiPhiloSofia** project aims at extracting, analyzing, and visualizing meaningful and interesting relationships among philosophers and philosophical concepts through the processing of the hyperlink structure and the textual and semantic content of Wikipedia pages. At the current stage of initial prototype implementation, we deal with 300 Western philosophers (or influential thinkers, writers, scientists, etc.), whose names we have extracted from Wikipedia's Timeline of Western Philosophers page.

Figure 1. Homepage of the WikiPhiloSofia portal.

Upon entering the portal, the user will choose to focus on one philosopher, two philosophers, or all 300 philosophers. In case a user chooses to focus on one philosopher, for example, the user is directed to the menu shown in Figure 2, which contains three fields corresponding to Subject, Predicate, and Object. The user can select a philosopher from the first dropdown menu, and then select a predicate from the second dropdown menu, in order to retrieve relevant objects. Table 3 summarizes various query options and result display modalities.

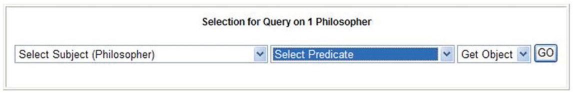

Selection for Query on 1 Philosopher
Select Subject (Philosopher)
Select Predicate
Get Object
GO

Figure 2. Menu for query on one philosopher.

Foci	Facets	Visualization Modalities
Facts and Links involving One Philosopher	Academic/Biographical Facts	Radial Graph View, Graph View, Tree View
	Direct Links/Influences	Radial Graph View, (Colored) Graph View
	Extended Links/Influences	Tree View, Radial Graph View, Graph View
Relations between Two Philosophers	Direct Relations	Radial Graph View
	Commonalities	Radial Graph View
	Direct (Common) Links/Influences	Radial Graph View
Networks and Rankings of All Philosophers	Strongest Link Networks	Graph View
	Strongest Influence Networks	Graph View
	(Purely Statistical) Rankings	Tag Cloud View

Table 3. Options for semantics-based search and exploration.

Interactive Visualization

The query results are presented via interactive visualization, implemented by using the Prefuse information visualization toolkit (http://prefuse.org).

Figure 3 presents a radial graph representing the facts about Plato, which amounts to visualizing the semantic network involving the philosopher. Figure 4 shows a fisheye tree representing extended influences originating from Plato.

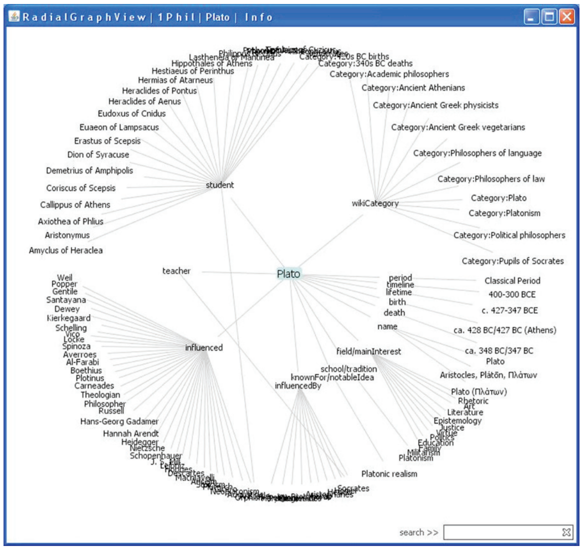


Figure 3. Academic/biographical facts on Plato.

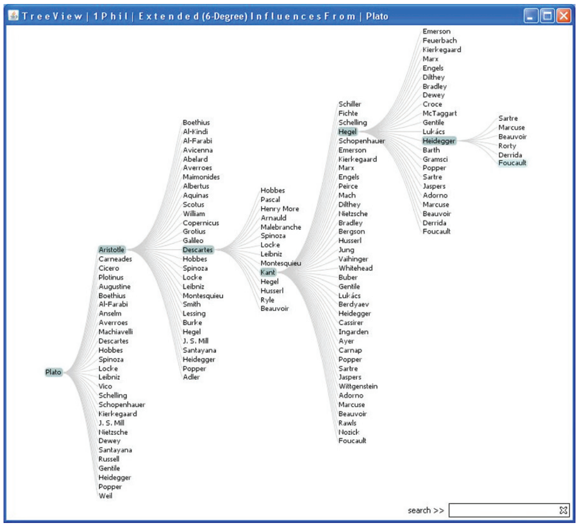


Figure 4. Extended influences originating from Plato.

Figure 5 shows a radial graph representing commonalities between Heidegger and Dewey.

The visualization of non-overlapping extended link/influence relations from/to one philosopher (using a graph or radial graph), and of strongest link/influence connections among 300 philosophers, is implemented by using a novel graph simplification method that we have developed, called the Strongest Link Paths (SLP) (Athenikos and Lin, 2008), which substantially simplifies the graph topology while highlighting the most dominant nodes and their interconnections.

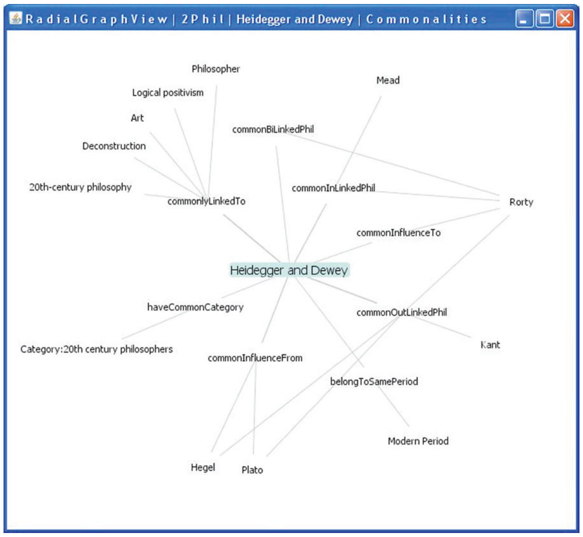


Figure 5. Commonalities between Heidegger and Dewey.

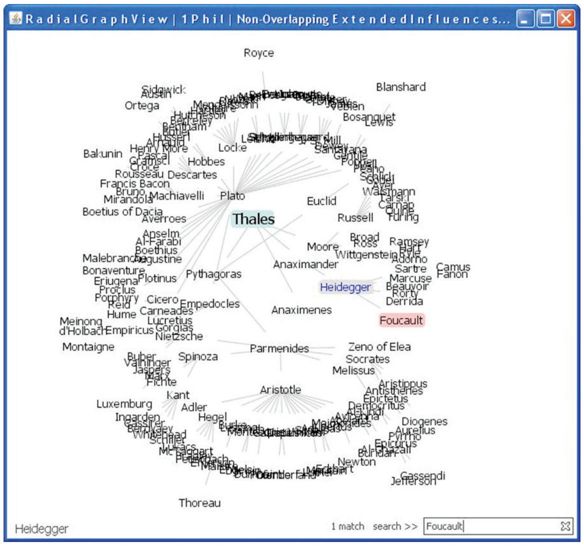


Figure 6. Non-overlapping extended influences originating from Thales.

Figure 6 presents a radial graph, simplified via SLP, which represents extended influences originating from Thales. This amounts to visualizing the small-world network (Milgram, 1967) of influence involving Thales. The figure shows that Thales, the first philosopher on the chronological list of 300 philosophers, can reach Foucault, the last one, within 3 degrees of separation (via Anaximander and Heidegger).

The graphs that result from applying SLP to the hyperlink/influence connections consist of distinct clusters separated from one another. Figure 7 shows a close-up of the largest cluster in the strongest out-link network that centers on Plato and Aristotle. Figure 8 shows the largest cluster in the strongest influenced-relation network, centering on Kant.

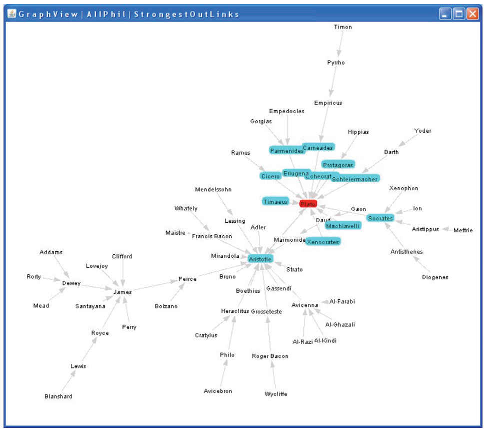


Figure 7. Largest cluster in the strongest out-link network.

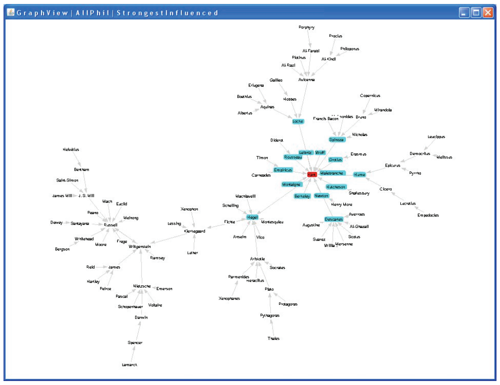


Figure 8. Largest cluster in the strongest influenced-relation network.

Comparison with AHCI Data

In this section we discuss some of the results of comparing the Wikipedia dataset against a subset of Thomson Reuters' Arts & Humanities Citation Index (AHCI) that contains 1.26 million records covering the 10-year period of 1988-1997.

Table 4 lists top 20 philosophers that receive the greatest number of in-links from among 300 philosophers in the Wikipedia dataset, those that receive in-links from the greatest number of philosophers in the Wikipedia dataset, and those (among the 300 philosophers) that have the highest citation count in the AHCI dataset.

Interestingly, Aristotle shows up on top for all 3 categories. While there are certain differences among the 3 lists, the majority of the philosopher names that appear on the 2 lists involving the Wikipedia dataset include major figures in the philosophy domain, as does the list obtained from the AHCI dataset. This shows that the hyperlink data extracted from Wikipedia, which embodies a huge amount of latent human annotation (Chakrabarti et al., 1999), provide a fairly good representation of the central figures in philosophy. In order to prevent a naïve, simplistic, and literal interpretation of these findings, it must be mentioned that what we argue is *not* that these philosophers are central figures *because* they have a large number of hyperlink connections in Wikipedia or that their relative centrality corresponds to link counts.

# of In-Links from Phil	# of In-Linked Phil	Citation Count
Aristotle (212)	Aristotle (105)	Aristotle (11466)
Plato (161)	Kant (89)	Freud (11455)
Kant (147)	Plato (88)	Foucault (11405)
Aquinas (128)	Leibniz (62)	Plato (10231)
Russell (109)	Hume, Locke (56)	Derrida (9822)
Hegel (98)	Russell (54)	Heidegger (8830)
Locke (95)	Spinoza (52)	Kant (8595)
Hume (92)	Hegel, Hobbes, Newton (47)	Cicero (8507)
Leibniz (87)	Descartes (46)	Nietzsche (6776)
Augustine (82)	Rousseau (43)	Marx (6681)
Descartes (79)	Augustine, Aquinas (42)	Hegel (5746)
Spinoza (75)	Bentham, Averroes (40)	Augustine (5471)
Avicenna (74)	Avicenna (39)	Wittgenstein (4767)
Hobbes (70)	Adam Smith (38)	Aquinas (4130)
Newton (69)	Voltaire (37)	Rorty (3630)
Averroes (66)	Nietzsche, Socrates, Democritus (36)	Weber (3600)
Marx, Albertus Magnus (64)	Marx (35)	Adorno (3594)
Nietzsche, Socrates, Democritus, Anselm (61)	Berkeley, Edmund Burke (33)	Sartre (3405)
Rousseau, Bentham, Adam Smith (60)	John Stuart Mill, William of Ockham, Grotius, Pierre Bayle, Thomas Reid, Al-Kindi, Al-Farabi, Al-Ghazali (32)	Rousseau (3070)
Bonaventure (59)	Hutcheson, Boethius, Wollstonecraft (31)	Descartes (2895)

Table 4. Highly in-linked vs. highly cited philosophers.

Table 5 shows a comparison of the list of philosophers who have bi-directional link connections with Heidegger in the Wikipedia dataset and the list of philosophers (considering only those that belong to the 300 philosopher set) that are most frequently co-cited with Heidegger within the AHCI dataset. As shown, 13 out of 20 most co-cited philosophers appear on the list of bi-linked philosophers. In addition, 3 out of the remaining 7 co-cited philosophers (Descartes, Wittgenstein, and Augustine) have hyperlink connections with Heidegger in one direction. Again, we are *not* making a naïve argument that those who are bi-linked or co-cited with Heidegger are intellectually closer or even similar to him in the order of bi-link/co-citation counts. It is however interesting to note the overlap between the two lists. In most cases shown in the table it is not hard to imagine why a certain philosopher may be bi-linked and/or co-cited with Heidegger, even though the reasons vary among the cases.

Bi-Linked Phil	Co-Cited Phil
Husserl (11)	Derrida (776)
Plato, Sartre, Derrida (8)	Kant (552)
Kierkegaard, Marcuse, Rorty (6)	Nietzsche (525)
Parmenides, Heraclitus, Aristotle, Nietzsche (5)	Hegel (453)
Anaximander, Brentano, Dilthey, Jaspers, Foucault (4)	Aristotle (412)
Duns Scotus, Schelling, Hartmann (3)	Husserl (407)
Kant, Hegel, Adorno (2)	Foucault (349)
	Plato (317)
	Wittgenstein (261)
	Rorty (258)
	Sartre (209)
	Adorno (205)
	Freud (193)
	Descartes (166)
	Marx (135)
	Kierkegaard (130)
	Jaspers (129)
	Aquinas (120)
	Lukács (104)
	Augustine (103)

Table 5. Philosophers bi-linked vs. co-cited with Heidegger.

Figure 9 shows a radial graph representing 25 philosophers most frequently co-cited with Heidegger and with Dewey, respectively. Those co-cited with Dewey include Heidegger. Those co-cited with both of them include Wittgenstein.

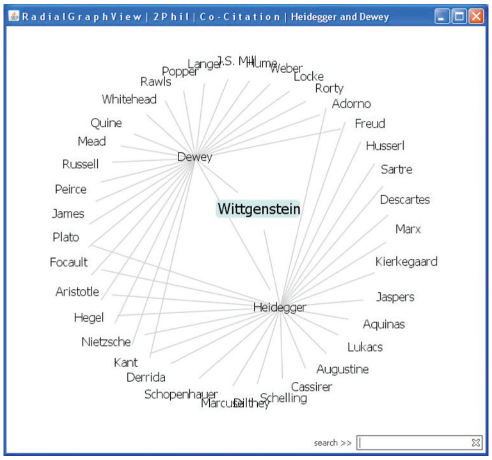

RadialGraphView | 2Phil | Co-Citation | Heidegger and Dewey

Figure 9. Philosophers most often co-cited with Heidegger and Dewey.

Related Work and Discussion

Wikipedia has recently become a topic of intense interest among researchers who recognize its utility as a source of a vast amount of knowledge that can be exploited for various applications. What renders Wikipedia a particularly valuable resource is the fact that it can be mined for knowledge based on its structural features as well as its textual content (Zesch *et al.*, 2007). As such, some Semantic Web (Berners-Lee *et al.*, 2001) researchers have turned to Wikipedia for clues to mitigating the knowledge acquisition bottleneck (Krötzsch *et al.*, 2005). In this paper we have demonstrated extracting/visualizing structured data available in Wikipedia by exploiting its hyperlinks (cf. Bellomi and Bonato, 2005), category links (cf. Chernov *et al*., 2006), and templates (i.e., infoboxes and wikitables) (cf. Auer and Lehmann, 2007). The logical extension of the current approach would be to extend the methodology to derive more extensive and implicit relations and connections, by exploiting the textual content of Wikipedia articles and by employing inference.

Social network analysis (SNA) has been used for some time in diverse disciplines besides sociology. With the advent of Web 2.0 (O'Reilly, 2005), characterized by the emergence of various collaborative authoring, blogging, bookmarking, tagging, networking, etc. sites that utilize combined social capital, SNA has become a key technique for capturing and exploiting data on social connections and interactions for various applications. Even though we have not attempted (and do not intend) to compute various centrality measures used in SNA (Wasserman and Faust, 1994), we have shown that the lists of highly-connected philosophers obtained by using even the simple hyperlink data *in general* provide a good representation of the central figures in the philosophy domain (*not* in the naïve sense that rankings purely based on hyperlink statistics correspond to the relative importance of each philosopher one-to-one). We have also shown that the networks of philosophers emerging from the hyperlink and semantic data extracted from Wikipedia exhibit the characteristic of the small world (Milgram, 1967) or the six degrees of separation phenomenon.

Insofar as the networks that we consider in this project are concerned with the connections among philosophers (and philosophical concepts), indirectly derived from Wikipedia, and not with the direct connections and interactions among the editors of the corresponding Wikipedia articles, the project is related to citation analysis, in particular, author co-citation analysis (ACA) (White, 2003). In this regard, we have presented a partial comparison of the results based on the Wikipedia data against those obtained by applying ACA to Thomson Reuters' AHCI data, using specific examples. While the results have shown an overall correspondence between the two datasets, it must be pointed out that the comparison was limited to 300 philosophers considered. It must again be emphasized that we do *not* equate link count or co-citation count with intellectual closeness or similarity.

Lastly, the WikiPhiloSofia project is prominently a project about visualization as an effective mode of data/information/knowledge representation. Information visualization, via the use of interactive, visual representations of abstract data, serves to amplify human cognition, making it possible or easier to recognize the hidden patterns and structures that might not otherwise be quite apparent or comprehensible (Card *et al*., 1997; Tufte, 1990), while enhancing the user's information search experience at the aesthetic level as well. We have illustrated how the results of various user queries can be presented using diverse modalities of visualization, by effectively visualizing the facts, relations, and networks that pertain to the 300 philosophers in the Wikipedia dataset. In particular, we have applied the Strongest Link Paths (SLP) method (Athenikos and Lin, 2008), which selects only the strongest link connections (measured in terms of hyperlink count or other connection strength measure) in order to highlight the most significant nodes and links. Even though SLP is rather simpler than other graph scaling methods such as pathfinder network (Schvaneveldt, Durso, and Dearholt, 1989) or main path analysis (Hummon and Doreian, 1989), we have found that it allows us to achieve substantial data reduction *and* to obtain a meaningful representation of the dominant figures and their connections within the network of philosophers even from the simple hyperlink data.

Conclusion

The WikiPhiloSofia project aims at creating a knowledge portal based on the data extracted from Wikipedia. In this paper we have illustrated extracting and visualizing the facts, relations, and networks involving 300 major philosophers in order to enable semantics-based search and exploration. The future work will include extending the approach to include more philosophers, extracting and visualizing the connections among philosophical concepts, deriving more extensive and implicit relations and connections, as well as applying the methodology to domains other than philosophy for comparison and evaluation purposes.

References

Athenikos, S.J. and Lin, X. (2008). The WikiPhil Portal: Visualizing Meaningful Philosophical Connections, Presented at 2008 Chicago Colloquium on Digital Humanities and Computer Science (DHCS 2008), Chicago, IL, November 2008. Forthcoming in *Proceedings of the Chicago Colloquia on Digital Humanities and Computer Science*.

Auer, S. and Lehmann, J. (2007). What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content, *Proceedings of 4th European Semantic Web Conference (ESWC 2007)*, Innsbruck, Austria, June 2007.

Bellomi, F. and Bonato, R. (2005). Network Analysis for Wikipedia, *Proceedings of the First International Wikimedia Conference (Wikimania 2005)*, Frankfurt am Main, Germany, August 2005.

Berners-Lee, T., Handler, J., and Ossila, O. (2001). The Semantic Web, *Scientific American*, 284: 34-43.

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (eds.). (1997). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufman Publishers, San Francisco, CA.

Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A. D., Gibson, D., and Kleinberg, J. (1999). Mining the Web's Link Structure, *Computer*, 32(8): 60-67.

Chernov, S., Iofciu, T., Nejdl, W., and Zhou, X. (2006). Extracting Semantic Relationships between Wikipedia Categories, *Proceedings of the First Workshop on Semantics Wikis (SemWiki 2006) at the Third European Semantic Web Conference (ESWC 2006)*, Budva, Montenegro, June 2006.

Hummon, N. P. and Doreian, P. (1989). Connectivity in a Citation Network: The Development of DNA Theory, *Social Networks*, 11: 39-63.

Krötzsch, M., Vrandečić, D., and Völkel, M. (2005). Wikipedia and the Semantic Web – the Missing Links, *Proceedings of the First Wikimedia Conference (Wikimania 2005)*, Frankfurt am Main, Germany, August 2005.

Milgram, S. (1967). The Small World Problem, *Psychology Today*, 1(1): 60–67.

O'Reilly, T. (2005). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (Last accessed 12 November 2008).

Schvaneveldt, R.W., Durso, F.T., and Dearholt, D.W. (1989). Network Structures in Proximity Data, *The Psychology of Learning and Motivation: Advances in Research and Theory*, vol. 24, G. Bower (ed.), 249-284. Academic Press, New York.

Tufte, E. R. (1990). *Envisioning Information*. Graphics Press, Cheshire, CT.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis. Methods and Applications*. Cambridge University Press, Cambridge, UK.

White, H. D. (2003). Pathfinder Networks and author Cocitation Analysis: A Remapping of Paradigmatic Information Scientists, *Journal of the American Society for Information Science and Technology*, 54(5): 423-434.

Zesch, T., Gurevych, I., and Mühlhäuser, M. (2007). Analyzing and Accessing Wikipedia as a Lexical Semantic Resource, *Proceedings of the Biannual Conference of the Society for Computational Linguistics and Language Technology*, Tübingen, Germany, April 2007.

BiblioMS: A Collaborative, Large-Scale Bibliography Management System

Neal Audenaert
Texas A&M University
neal@cs.tamu.edu

Richard Furuta
Texas A&M University
furuta@cs.tamu.edu

The need to create, maintain, and publish a bibliography has been a recurring theme in the digital humanities projects we have been involved with over the past fifteen years. This may be as simple as listing references for individual digital artifacts or as complex as maintaining a comprehensive record of all significant works published about a particular author. Bibliographies are rarely the main research focus of a project—from either the humanities or the technical side—and are therefore frequently relegated to the sidelines. In the best of scenarios, custom applications are built that support the bibliography of a particular project. In the worst (and perhaps more common) scenarios, a scholar creates a bibliography in Word and the result of the "export to HTML" function is copied into the appropriate page on the project's Web site.

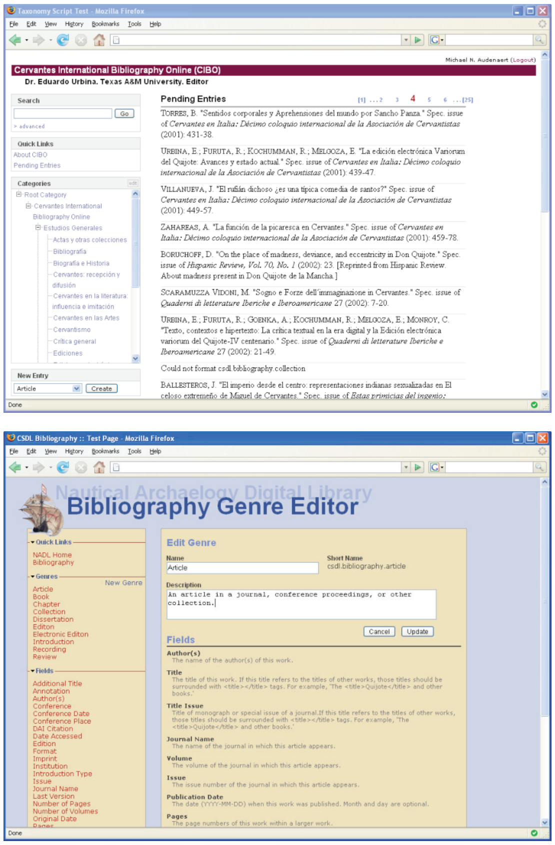
Whether the bibliographic component of a digital humanities project is a necessary chore or a key scholarly contribution, for most projects it is a tedious, labor-intensive effort. The development of one-off bibliography solutions for individual projects may reduce the effort required by humanities personnel, but this comes at the cost of increased developer time—a scarce resource. While developing a bibliography system from well-defined initial specifications is straightforward, it is tedious and time consuming. Furthermore, unless care is taken in the initial design, the resulting system is not likely to anticipate future needs of the project making it difficult or impossible to modify the application to meet new needs as they arise.

To avoid developing a new bibliographic management system (BMS) from scratch for each new project we work on, we are implementing a general-purpose system that can be tailored to meet project specific needs. Whereas the development of a BMS tailored to the specific needs is of a single project is straightforward, the development of a general-purpose BMS poses interesting challenges. We have identified four key requirements based on our experience with the use of bibliographies in various projects we have been involved with.

First, the BMS must support **user-defined genres**. While standard bibliography genres (such as books, journal articles, electronic resources) are usually adequate for personal use, editors in charge of large scale bibliography projects often need to tweak the standard genres in order to convey unique aspects of their material more effectively. Special purpose projects may need to record material not envisioned when the BMS was initially developed. In addition to defining the content of different genres of bibliographic entries, a BMS also needs to define semantic relationships between genres. For example, a chapter is a part of a book, or a review is a journal or newspaper article about another work. While these relationships are expressed in a human readable form in the entry's content, it is also necessary to provide formal, machine-readable representations in order to build systems that leverage these relationships to support data entry and browsing.

Second, the BMS must enable **multi-faceted organization and navigation** of the collection. This includes standard features such as full text searching and the ability to search or browse based on entry fields (a challenging task if those fields are not known in advance). In addition to these techniques, editors of large bibliographies often rely on hierarchical taxonomies and controlled vocabularies to structure collections. The BMS should provide tools to assist editors in developing these organizational strategies. While most projects we are familiar with rely on a unified categorization scheme developed from a single, authoritative editorial perspective, we anticipate the need to organize collections from multiple perspectives. One example of this is allowing individual users to add their own tags or to bookmark entries in "folders." By allowing these personal organizations to be either private or public, a BMS can enable external editors to build on existing work to offer alternative views of the bibliographic record of a field.

Third, the BMS must facilitate **collaborative editing**. In a typical, project-scale bibliography, a single editor or team of editors is responsible for maintaining the quality and accuracy of the bibliographic entries. Much of the day-to-day work of entering data and updating entries, however, may be performed by assistants, often graduate students working under the supervision of the editor (or editorial board). Moreover, the broader academic community may assist this team of editors and assistant editors by recommending new items for inclusion in the bibliography and suggesting corrections to existing items. To support this workflow, the BMS needs to provide a sandbox where users with lower privileges may edit entries prior to being made publicly available by the approval of an authorized editor. This also requires the ability to assign users to different editorial access levels, to track changes, and to maintain persistent links to the different published states of a bibliographic record. The BMS should also include user management features to allow users to maintain their own profile and to allow project editors to grant and revoke editorial permissions.


Taxonomy Script Test - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Michael N. Audenaert (Logout)
Cervantes International Bibliography Online (CIBO)
Dr. Eduardo Urbina. Texas A&M University. Editor
Search
Go
> advanced
Quick Links
About CIBO
Pending Entries
Categories
edit
Root Category
Cervantes International Bibliography Online
Estudios Generales
Actas y otras colecciones
Bibliografía
Biografía e Historia
Cervantes: recepción y difusión
Cervantes en la literatura: influencia e imitación
Cervantes en las Artes
Cervantismo
Crítica general
Ediciones
New Entry
Article
Create
Pending Entries
[1] ...2 3 4 5 6 ...[25]
Done

TORRES, B. "Sentidos corporales y Aprehensiones del mundo por Sancho Panza." Spec. issue of *Cervantes en Italia: Décimo coloquio internacional de la Asociación de Cervantistas* (2001): 431-38.

URBINA, E.; FURUTA, R.; KOCHUMMAN, R.; MELGOZA, E. "La edición electrónica Variorum del Quijote: Avances y estado actual." Spec. issue of *Cervantes en Italia: Décimo coloquio internacional de la Asociación de Cervantistas* (2001): 439-47.

VILLANUEVA, J. "El rufián dichoso ¿es una típica comedia de santos?" Spec. issue of *Cervantes en Italia: Décimo coloquio internacional de la Asociación de Cervantistas* (2001): 449-57.

ZAHAREAS, A. "La función de la picaresca en Cervantes." Spec. issue of *Cervantes en Italia: Décimo coloquio internacional de la Asociación de Cervantistas* (2001): 459-78.

BORUCHOFF, D. "On the place of madness, deviance, and eccentricity in Don Quijote." Spec. issue of *Hispanic Review, Vol. 70, No. 1* (2002): 23. [Reprinted from Hispanic Review. About madness present in Don Quijote de la Mancha.]

SCARAMUZZA VIDONI, M. "Sogno e Forze dell´immaginazione in Cervantes." Spec. issue of *Quaderni di letterature Iberiche e Iberoamericane* 27 (2002): 7-20.

URBINA, E.; FURUTA, R.; GOENKA, A.; KOCHUMMAN, R.; MELGOZA, E.; MONROY, C. "Texto, contextos e hipertexto: La crítica textual en la era digital y la Edición electrónica variorum del Quijote-IV centenario." Spec. issue of *Quaderni di letterature Iberiche e Iberoamericane* 27 (2002): 21-49.

Could not format csdl.bibliography.collection

BALLESTEROS, J. "El imperio desde el centro: representaciones indianas sexualizadas en El celoso extremeño de Miguel de Cervantes." Spec. issue of *Estas primicias del ingenio:*

Edit Genre

Name: Article

Short Name: csdl.bibliography.article

Description: An article in a journal, conference proceedings, or other collection.

Fields

Author(s)
The name of the author(s) of this work.

Title
The title of this work. If this title refers to the titles of other works, those titles should be surrounded with <title></title> tags. For example, 'The <title>Quijote</title> and other books.'

Title Issue
Title of monograph or special issue of a journal.If this title refers to the titles of other works, those titles should be surrounded with <title></title> tags. For example, 'The <title>Quijote</title> and other books.'

Journal Name
The name of the journal in which this article appears.

Volume
The volume of the journal in which this article appears.

Issue
The issue number of the journal in which this article appears.

Publication Date
The date (YYYY-MM-DD) when this work was published. Month and day are optional.

Pages
The page numbers of this work within a larger work.

Fig 1: Screen captures of the main browsing interface deployed for the Cervantes Project and the editing tools for bibliographic genres deployed for the Nautical Archaeology Digital Library

Finally, the BMS must allow **integrated access** with the project's existing Web-interfaces. This includes developing editors' and readers' interfaces whose look and feel match that of the rest of the project's Web site. Components of the bibliography should be accessible throughout the project site. For example, it should be possible (and relatively easy) to include a reference to a bibliography entry in the text that appears on the site or in a reference section for a particular digital artifact. It should also be possible to integrate bibliography editing tools with other editorial interfaces. For example, if a nautical archaeology project is building an interface to allow project members to add information about a shipwreck, it should be possible to add references to supporting material directly from this interface without going to a separate editor's interface for the BMS.

Before deciding to develop our own BMS, we first investigated the many open source and commercial solutions currently available including BibTeX, Zotero, and EndNote. Like others (Stout 2008), we quickly concluded that most of these systems are tailored to the management of personal collections and inadequate for our purposes. RefDB provides reference management engine that might be useful, but would require significant customization and extension to meet our needs. Given the amount of work that would be required to extend any of these existing systems, we decided that developing our own BMS was the best course of action.

Implementation

To achieve this, we have developed a system called BiblioMS. Instead of implementing BiblioMS as a standalone Web application with it own built-in interface, we have designed it to function as a set of application modules that can be integrated with existing Web sites or other applications. This allows us to maintain the distinctive look-and-feel of the different projects in which we deploy BiblioMS. For example, Fig. 1 shows screen captures from two projects in which we have deployed this system.

The main BiblioMS system consists of the core bibliography management engine and a set of JavaScript libraries for accessing this engine. The management engine implements the storage, modification, revision control, categorization, and retrieval requirements outlined above. A Web interface layer provides access to the engine via an HTTP API (application programming interface). This architecture is shown in Fig 2.

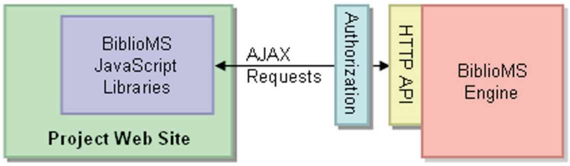


Fig 2: BiblioMS architecture

The JavaScript libraries implement an object-oriented API that allows project developers to interact with the management engine directly from Web pages without writing any server side code. In our experience, we have found that developing user interfaces in HTML and JavaScript and using server-side technologies (such as PHP and Java Servlets) only for data management improves the modularity of our tools and makes our applications easier to maintain and modify.

The BiblioMS engine itself is aware of users (in order to record who is making changes to entries and to maintain personal annotations and collections) but does not provide user authentication and authorization services. Instead, an authentication layer provides these services by filtering access to the HTTP interface. The advantage of this approach is that administrators can configure and control user access on a feature-by-feature basis by editing an XML configuration file. In our current projects, we have used a three level permissions strategy (editors, assistant editors, and external contributors) but this implementation allows alternative approaches.

Conclusions

Our work offers two primary contributions. First, based on our experience with numerous digital humanities projects, we have described four key requirements for developing a general-purpose bibliography management system: user-defined genres, multi-faceted organization and navigation, collaborative editing, and integrated access. Second, we have demonstrated our implementation of a system that meets those requirements. We plan to build on our current work by using this system as a basis for studying how large, project-scale bibliographies are used both by their editors as well as by their readers. We intend to use the results of these studies to improve finding aids, streamline editorial workflow, improve quality control mechanisms, and support user-centric organization and annotation of the collection.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0534314.

References

[BibTeX] BibTeX. http://www.bibtex.org/ (13 November 2008)

[EndNote] EndNote. http://www.endnote.com/ (13 November 2008)

[RefDB] RefDB. http://refdb.sourceforge.net/ (13 November 2008)

[Stout 2008] Stout, J., Wulfman, C., and Mylonas, E. A bibliographic utility for digital humanities projects. In Proceedings of Digital Humanities 2008, Oulu, Finland, 24-29 June, 2008. Univesity of Oulu.

[Zotero] http://www.zotero.org/ (13 November 2008)

The LANCHART Search Engine—Making important progress in data and data archiving reuse

Michael Barner-Rasmussen
University of Copenhagen
mbr@hum.ku.dk

In this paper, we describe the basic design of a multi purpose, multi disciplinary database-driven research tool for spoken language research developed and implemented by the LANCHART Centre (University of Copenhagen). The LANCHART Search Engine provides an open source, highly robust, fast and versatile framework for the entering of, working with/analyzing, visualizing, and not least reusing spoken language research resources.

Secondly, we describe the IS-structures and programs supporting the search engine, most notably the creation of import-export filters for a number of widely used transcription and analysis tools (presently Praat, Transcriber, CLAN).

Last, to illustrate the concrete usability and versatility of the resultant e-science tool, we demonstrate by example some of the many uses this e-science tool has already been put to use by LANCHART as well as by visiting scholars.

Background

The LANCHART project (Gregersen 2009, lanchart.dk) initially created the search engine for contextualized searching in/mining of a very large spoken language corpora presently consisting of ~380 sociolinguistic interviews spanning more than 30 years of same-respondent recordings. During the course of its development it has proven itself an obvious candidate to an e-science tool for other corpora of spoken language under the Danish CLARIN project (http://english.dkclarin.ku.dk/). As it stands the corpus encompasses ~7 M transcribed words and ~39 M linguistic annotations.

The LANCHART search engine is a working research tool that is supporting more than 15 different thus computer assisted research efforts ranging from phonetic, through grammatical to discourse context analyses.

Architecturally this versatility, i.e. reuse of both research data and the underlying enabling technology, has been achieved by stratifying the basic data model into elements:

- topology or ‘data architecture’
- (analytical) content
- semantic structure(s)

This separation has long been advocated by prominent new media scientists such as Bolter (2001), Weinberger (2002), Levy (1998) and has proved rarely opportune in the case of preparing scientific work on spoken language corpora for an up to date e-science.

Data Architecture, outputs and the reusable engine

The data architecture is dictated solely by the structural elements of spoken language, i.e. it happens in contiguous time, there may be many simultaneous speakers and many things (of analytical interest) may happen at once.

These structural characteristics are afforded by designing the data store around two basic entities:

1. ‘Slice’: a period of time defined by its starting and stopping time relative to the recordings (recording may be audiovisual or audio only).
2. ‘Tier’: a set of slices spanning the length of the sample

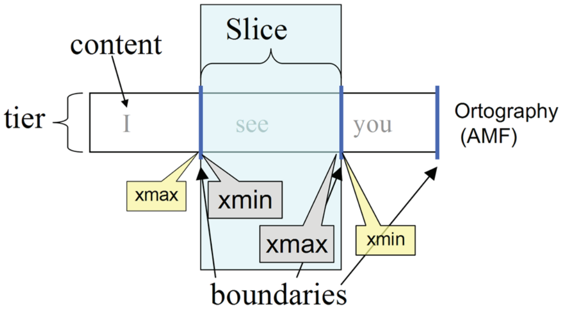


Fig 1: The relationship between slices and tiers, times and analytical content

Each slice may be annotated, e.g. with the word uttered or a comment, but only one such annotation may be entered for any one slice. Simultaneously occurring annotations are represented by placing a slice with identical boundaries in a different tier.

Tiers are identified by a unique tier name (Orthography (AMF) in Fig. 1.) allowing for linking to speaker metadata, annotation encoding convention, lists of related tiers etc.

This purposefully devoid-of-meaning data architecture creates an environment where most needs of researchers of spoken language are met. They may annotate whatever, wherever and importantly *whenever* (i.e. as finely grained) as they wish.

The search engine on its part has two output formats: a KWIC concordance of the finds and a .csv file in which the finds are supplemented with columns containing time values, the content of aligned intervals in all other tiers and the speakers’ background information. This is then used for statistical analyses using programs such as SAS.

Data Semantics and reuse of data

As hinted to above, the semantics, the meaning of the annotations in any given tier, is entirely decided by the annotator. The unique name of each tier allows for any amount of metadata to be linked with the tier, e.g. guides and explanations of the contents as well as coding conventions used.

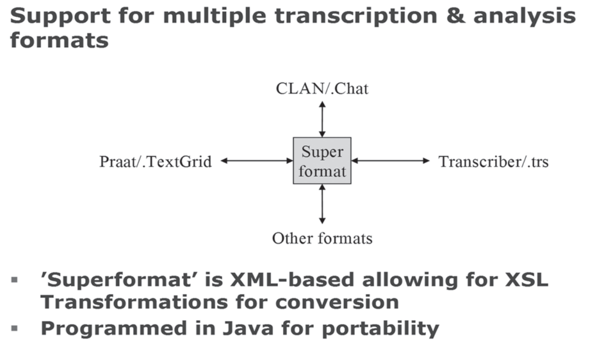
Reuse becomes the simple task of choosing between already created tiers: which are of analytical interest for the present purpose? Then create a ‘package’ with only those tiers present, export this package to one of several supported formats and get started on one’s own analytical contribution.

IS structures and computer programs supporting the search engine

A search engine is not much of a research tool if it isn’t easy to input data for searches, does not scale or if it is not well documented, easily customized and written in a easy to maintain programming language.

With this in mind from the very outset, the LANCHART Search Engine is developed in open source tools (LAMP, Java/jsp) with initial support for the CLAN, Praat & Transcriber spoken language representational formats.

It is also central to note, that the conversion and import functions were implemented by translating to and from a ‘super-format’ instead of directly between formats. This allows the researchers freedom to choose whatever tool suits their situation best and for easy, if not always lossless, translation between all formats used both currently and in the future – writing new translators is, if not easy, then thanks to the architecture surmountable.



Support for multiple transcription & analysis formats

- **'Superformat' is XML-based allowing for XSL Transformations for conversion**
- **Programmed in Java for portability**

Fig 2: The design scheme for conversion and translation between formats usable by the LANCHART Search Engine

Ongoing computer assisted analysis / e-science using the LANCHART Search Engine

Here, we mention but a few of the LANCHART Centre's corpus/e-science research projects (for more see http://lanchart.hum.ku.dk/reports):

In most detail, we describe the work of Jeffrey Parrot who has formulated a project investigating socially salient and theoretically challenging variation in the pronominal case forms of Danish within the broader context of research on the morphosyntax of case in Germanic, to test whether variation in Danish pronominal case is stable or represents a change in progress. By using the analytical tiers already present in the corpus as his research base he can then add his own tiers considering factors including attested pronouns' case form, person, number, conjunct ordering, and structural position as well as including speakers' age, sex, geographic location, profession, and relative literacy. Using the search engine to locate candidate concordance sections, Jeffrey may jump start this effort.

Other researchers at the centre have produced works on the correlation of phonetic and grammatical variables, on bilingualism, and perhaps most notably, phonetic variation in real time over recordings spanning more than 30 years (see Gregersen 2009).

The digital infrastructure makes it possible not only to study the influence of geographical origin, gender and social class on the use of the linguistic variables in question and the possible interaction of functional and sociolinguistic factors but also to analyse possible correlation between the different linguistic variables and between linguistic variables and discourse context allowing us to formulate and test hypotheses about the origin and spreading of patterns of linguistic changes in late 20th century Danish.

References

Blanke, Tobias et. al. (2008). e-Science in the Arts and Humanities – A methodological perspective in Opas-Hänninen, Lisa Lena (Ed.) et. al.(2008): Digital Humanities 2008. Book of Abstracts.

Bolter, Jay David (2001). Writing Space: Computers, Hypertext, and the Remediation of Print, Second Edition. Mahwah, NJ: Lawrence Erlbaum.

Juel Jensen, Torben (2009): "Generic variation? Developments in use of generic pronouns in late 20th century spoken Danish" In Acta Linguistica Hafniensia Vol. 41.

Levy, Pierre (1998). Becoming Virtual: Reality in the Digital Age. Da Capo Press.

Normann Jørgensen, Jens (2005): "Gender differences in the development of language choice and patterns in the Køge Project" in Code-Switching in the Køge Project. Special issue of the *International Journal of Bilingualism* 7:4, 2004 (publ. 2005)

Weinberger, David (2002). Small Pieces Loosely Joined: A Unified Theory of the Web. NY: Perseus Publishing

Gregersen, Frans. Ed.(2009): Acta Linguistica Hafniensia Vol. 41. Language Change In Real Time. Evidence from the Danish Laboratory. *International Journal Of Linguistics* (The Linguistic Circle of Copenhagen)

In the Header, but Where?

Syd Bauman
Brown University
Syd_Bauman@Brown.edu

Dorothy Carr Porter
Digital Humanities Observatory
dot.porter@gmail.com

The TEI header is a valuable but sometimes overlooked part of a TEI document. The header is the main source of documentation for a TEI encoded electronic document, and has been created to describe "an encoded work so that the text itself, its source, its encoding, and its revisions are all thoroughly documented."[i] It is described as an electronic analogue to the title page of a printed book, but it is really much more than that. The header provides a location to place any kind of information about a text that should not or need not be described within the body of the text itself—for example, narrative descriptions of a series of illustrations, or of individual illustrations; definitions of terms that occur in the text; demographic information about the people involved in the creation of the text, its TEI transcription, or mentioned in the text; even complete descriptions of physical objects (using the <msDescription> element). All these can be placed in the header and then linked to passages in the body of the text using one or more of a variety of available linking mechanisms. This combination of digital textual data (stored in <text>) and digital header metadata (stored in <teiHeader>) provides a method for description that is generally clear and flexible. Witness that several other popular XML vocabularies use an analogous system of separating metadata from data immediately as the two children of the root node, with the metadata coming first. (Notably XHTML[ii] and DocBook[iii] follow this pattern.) However, the placement of certain specific metadata within the header is not always so clear.

The TEI *Guidelines* often recommend that users place a specific piece or kind of metadata in the TEI header, but sometimes they do not specify into which area of the header the particular information should be placed. E.g., the <gap> element "indicates a point where material has been omitted in a transcription, whether for editorial reasons described in the TEI header, as part of sampling practice, or because...". However, a search through the chapter on the TEI header does not clearly suggest where exactly those editorial reasons should be described (although there is an example of <samplingDecl> which does explain the use of <gap> for editorial reasons).

There are some 218 references to the TEI header in the source of the TEI *Guidelines* that are neither within the chapter that directly discusses the TEI header, nor are within an example, a heading, or the <teiHeader> of the TEI Guidelines themselves. In how many of these instances do the Guidelines make clear recommendations? And when there are no clear guidelines, how do users decide how to use the header?

In our paper, we will answer the question "in how many places do the TEI *Guidelines* recommend placing information in the TEI header, but do not specify exactly where?". Furthermore, we hope to examine each such occurrence in some more detail, even recommending where such information be placed for as many occurrences as possible.

For some such occurrences, we also hope to ask a set of TEI users where they are placing this information, or, if they do not happen to need to record this particular type of metadata, where they would be inclined to place it if they did.

Finally, based on the research conducted, we hope to be able to come to some broad conclusions as to how much of a problem this issue actually represents.

We will first study the *Guidelines* to find all the instances where users are told to use the header, but are not told exactly how. We will then conduct at least informal conversations with TEI users (both long-term users and novices), if not use a formal survey instrument, to try to ascertain how they do (or would be inclined to) act in the given situations. We will discuss what impact the solutions users come up with might have on both interchange and interoperability. We will also attempt to ascertain how much of a perceived problem this issue really presents for these users. Our own experience with the TEI, as users and developers, will also be helpful for this study.

Finally, for as many of the cases for which the TEI Guidelines do not make a specific recommendation as possible, we will attempt to provide a suggestion for precisely where in the TEI header the information should be placed. We do this with the hope that it will both improve the interchange the TEI supports, and lower the barrier to adoption of TEI by new projects.

[i] Burnard, Lou and Syd Bauman, eds. "The TEI Header" *Guidelines for Electronic Text Encoding and Interchange*. 2007-11-01. http://www.tei-c.org/release/doc/

tei-p5¬doc/html/HD.html (24 Nov 2007).

[ii]"XHTML" *The Extensible HyperText Markup Language* (Second Edition). 2002-08-01. http://www.w3.org/TR/xhtml1/ (24 Nov 2007).

[iii]Walsh, Norm and Leonard Muellner. *DocBook, The Definitive Guide*. O'Reilly, 1999. urn:isbn:1-56592-580-7.

Automatisches Metrisches Markup

Klemens Bobenhausen
Universtität Freiburg
klemens.bobenhausen@germanistik.uni-freiburg.de

Gedichte mit metrischen Annotationen zu versehen ist eine sehr zeitaufwändige Angelegenheit. Für das Projekt „Freiburger Anthologie - Lyrik und Lied" habe ich Hunderte von Gedichten manuell metrisiert.[1] Aus dieser Arbeit heraus ist die Idee erwachsen, eine automatisierte Technik zu entwickeln, schriftlich niedergelegte deutschsprachige Gedichte automatisch metrisieren zu lassen und sie in ein TEI P5-kompatibles XML umzuwandeln.[2]

Nach gut einem Jahr Arbeit kann ich nun ein sehr passa-bles Ergebnis präsentieren. Deutschsprachige Gedichte lassen sich zu einem sehr hohen Prozentsatz automatisch metrisch bestimmen. Je nach der Anzahl an Fremdwörtern, die ein Gedicht aufweist, schwankt die Qualität zwischen 90% und 100%. Ein Text ohne Fremdwörter wird durchweg zu einem Prozentsatz von 99% bis 100% richtig erkannt.[3]

Um dieses Ergebnis zu erreichen, wird der Text zunächst automatisch in seinen metrisch relevanten Strukturen erkannt. Dazu gehört die Erkennung der einzelnen Strophen, der Verse, der Wörter und der Silben. Für die Silbensegmentierung haben wir ein eigenes System entwickelt. Bis dieses ausgereift ist, verwenden wir in der Entwicklungsumgebung noch eine sehr gut funktionie-rende, aber noch unveröffentlichte Beta-Version einer orthographisch basierten Silbentrennung aus dem Projekt „Segmenti".[4]

Nach der Strukturerkennung des Textes werden sogenannte „prosodische Vorhersagen" der jeweiligen im Text vorkommenden Wörter erstellt. Dabei greifen wir auf eine Vielzahl an definierten Regeln zurück. Das ges-amte Regel-Set umfasst derzeit 35 Regeln, einige davon sind Eigenentwicklungen, andere wurden aus den Erkenntnissen unterschiedlichster germanistischer Disziplinen übernommen oder abgeleitet. Das Regel-Set reicht vom „deutschen Initialakzent" über „offene Tonsilben" und die „Penultimaregel" (die allesamt für die gesamte deutsche Sprache gelten) bis hin zu gedichttypischen Erscheinungsformen wie z.B. die Betonung von einsilbigen Wörtern am Versende. Grundlegend unterscheiden wir zwischen „prosodischen

Vorhersagen“, die für eine Betonung einer Silbe sprechen – und Vorhersagen, die eine unbetonte Silbe wahrscheinlich machen.

Alle diese Regeln werden automatisiert angewendet und für jede Silbe gesammelt. Das Ergebnis sieht dann wie folgt aus. Ein „x“ steht für die erste Silbe eines Wortes, die nach dem Regeln des germanischen Initialakzentes (auch foot form trochaic rule) für eine Betonung der Silbe spricht. Ein oder mehrere „+“ für die Wahrscheinlichkeit, dass es sich bei der Silbe um eine betonte Silbe handelt. Ein „-“ steht für den Verdacht, dass es sich bei der Silbe um eine unbetonte Silbe handelt. Eine „0“ für eine Silbe, für die keine Verdachtsmomente vorliegen. Für jede Silbe wird die Kette der Regeln gespeichert.

Fried | lich be | kämp | fen
Nacht sich und Tag.
Wie das zu dämp | fen,
Wie das zu lö | sen ver | mag![5]

Fried =x000000+0000000000000000000000
0++000000000 =x+++

lich =0000000000000000000000000-00000-
0000000000 =--

be =x000000000000000000000000000000000000
00000 =x

kämp =0+00000+00+0000000++0000000000000000
000000 =+++++

fen =00000000000000000000000000-
000000000000000 =-

Im Anschluss daran wird eine Euphonie-Berechnung durchgeführt, die von der Idee geleitet wird, dass mit hoher Wahrscheinlichkeit innerhalb eines Verses keine zwei betonten Silben nebeneinander liegen (Ausnahmen des Hebungspralls hier außen vor gelassen) – und keine drei unbetonten Silben aufeinander folgen dürfen. So wird z.B. das „x“ auf der Silbe „be“ unbetont, da die benachbarte Silbe „kämp“ durch ihre massiven Betonungshärte die benachbarte Silbe überschreibt. Die Euphonie-Berechnung erfolgt ebenfalls vollautomatisch und ist in einem Algorithmus festgehalten (+ steht für eine Silbe, die den Verdacht einer Betonung erhält, - für eine Silbe, die den Verdacht einer Unbetonung erhält).

Nach der Euphonie-Berechnung sieht das entsprechende Schema dann wie folgt aus:

Fried| +	lich -	be| -	kämp| +	fen -		
Nacht 0	sich 0	und -	Tag. 0			
Wie 0	das -	zu -	dämp| +	fen, -		
Wie 0	das -	zu -	lö| +	sen -	ver| -	mag +

Im anschließenden Analogieverfahren wird statistisch berechnet, wie die noch nicht bestimmten Silben betont sein könnten. Dabei wird z.B. geprüft, ob eine feste Anzahl von betonten Silben für jeden Vers vorliegt, die auch für die anderen Verse angenommen werden kann, ob über die zuvor erkannte Reimstruktur am Ende des Verses unbestimmte Silben bestimmt werden können (da zwei sich reimende Verse im Bereich des Reims eine identische Betonungsstruktur aufweisen müssen). Die wichtigeste Analogie-Regel ist jedoch darüber definiert, dass noch nicht erkannte Silben mit hoher Wahrscheinlichkeit die gleiche Betonung aufweisen wie Silben anderer Verse auf der selben Position mit gleicher Silbenzahl. Am Ende der Analogie-Berechnung werden die Zeichen + und - (die bislang prosodische Betonung angezeigt haben) in metrische Zeichen (in der Bedeutung von Hebungen und Senkungen) umgewandelt. Im besten Fall liegt dann das komplette metrische Schema der Strophe vor:

Fried| +	lich -	be| -	kamp| +	fen -		
Nacht +	sich -	und -	Tag. +			
Wie +	das -	zu -	damp| +	fen, -		
Wie +	das -	zu -	lo| +	sen -	ver| -	mag +

Das Ergebnis wird nun mit einer umfangreichen Datenbank verglichen, in der aufgrund der Standardwerke zur metrischen Bestimmung deutscher Strophenformen an die 4500 unterschiedliche Strophenformen versammelt wurden.[6]

In etlichen Fällen lässt sich über alle Strophen hinweg eine gleichbleibende Struktur erkennen, die jedoch an einzelnen Stellen unterbrochen scheint. In seltenen Fällen liegt an solchen Stellen eine metrische Fehlinterpretation des Programms zu Grunde, viel häufiger jedoch ist der Fall, dass der Autor des Textes die allgemeinsprachlich gültige Prosodie eines Wortes in ein Reibungsverhältnis zur metrischen Struktur setzt, wie z. B. in dem Wort „nachlässiger“ im Gedicht „Die Beiden“ von Hugo von Hofmannsthal. Im Prinzip lassen sich solche Stellen automatisiert erkennen und „schwebende Betonungen“, „Emphase“ und prosodisch neutrale oder schwache Stellen in einem Gedicht nachweisen.

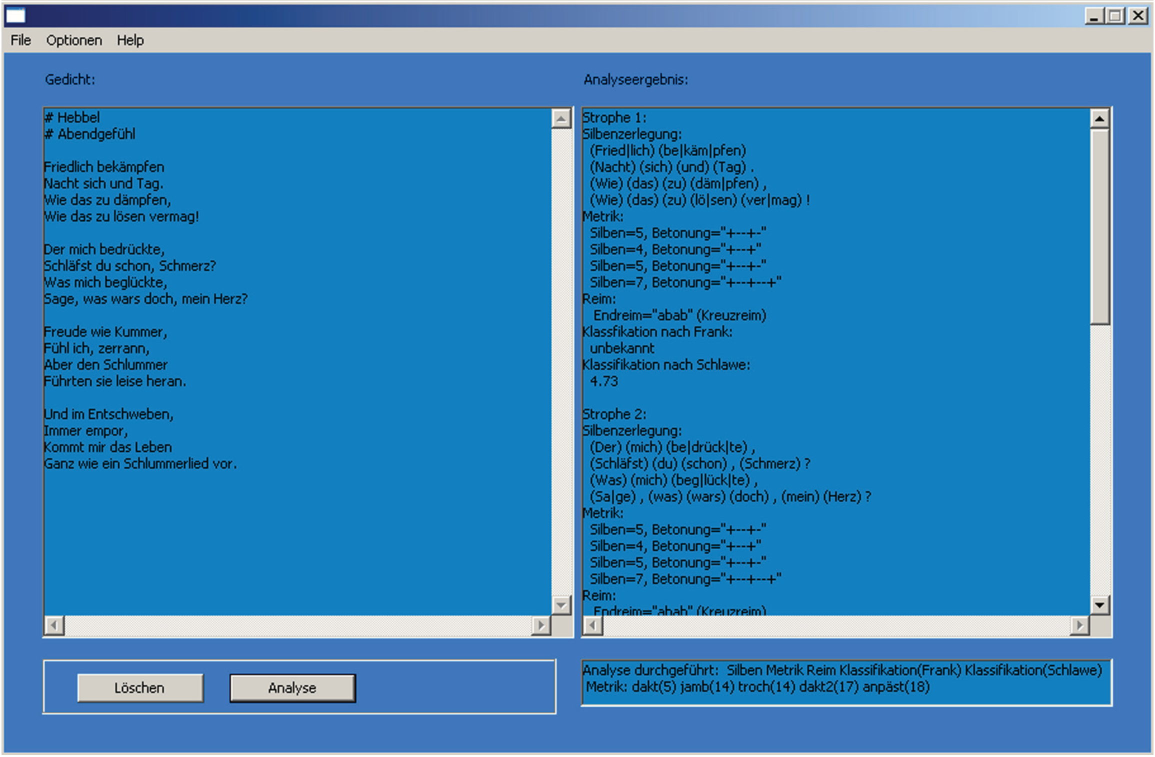
Die vorgestellte Technik ist derzeit noch nicht endgültig programmiert. Derzeit kann der „metricalizer“[7] mit einem sehr kleinen Regel-Set von zwei Regeln nur „regelmäßig“ metrisierte Gedichte erkennen.[8] Das gesamte Regel-Set liegt derzeit nur in Form von Macros vor, die als Kette in der Textverarbeitung „Word Perfect“ ausgelöst werden können. Bis auf die Silbensegmentierung und die Analogieberechnung ist das System – obwohl nur in einer Textverarbeitungssoftware – bereits voll automatisiert.

Aus den erzeugten Ergebnissen lässt sich eine TEI P5-konforme Struktur des Textes erzeugen, in der bis auf die einzelne Silbe hinab die metrische Struktur annotiert werden kann.

Mit der aufgezeigten Technik ist es möglich, Gedichte der letzten vier Jahrhunderte metrisch zu erkennen. Da wir nicht auf Wörterbücher zurückgreifen, in denen die Betonungswerte der einzelnen Wörter vordefiniert wird, ist es zudem möglich, auch Texte zu analysieren, die nicht der heute gültigen Standard-Orthographie folgen.

Für die Sprach- und Literaturwissenschaft bietet der Ansatz eine Reihe von interessanten Möglichkeiten, die von der automatisierten Metrisierung ganzer Korpora über die Text-to-Speech-Forschung reichen, woraus sich entsprechende Forschungs- und Unterrichtsansätze konstruieren lassen. Im Prinzip ist die angewendete Technik nämlich nicht nur in der Lage, Gedichte zu metrisieren, sondern ganz allgemein prosodische Vorhersagen für jeden gedruckten Text zu erstellen. Das Fernziel der Unternehmung ist es jedoch, die automatische metrische Bestimmung von Gedichten mit der automatisierten Textgenerierung zu verknüpfen, um den Rechner eines Tages in die Lage zu versetzen, metrisch vollendete Gedichte selbständig zu produzieren.

Auf der Tagung werde ich kurz die nationalen und internationalen Vorgängerprojekte benennen, auf die wir uns beziehen können, die logische Seite des Projektes vorstellen (vor allem die Regeln der prosodischen Vorhersage) und auf die theoretischen Möglichkeiten und Schwachstellen (Fremdworterkennung, Ketten von einsilbigen Wörtern) eingehen und anhand eines Ad-hoc-Beispiels aus dem Publikum den Beweis für die Funktionalität des Systems erbringen.[9]



```
# Hebbel
# Abendgefühl

Friedlich bekämpfen
Nacht sich und Tag.
Wie das zu dämpfen,
Wie das zu lösen vermag!

Der mich bedrückte,
Schläfst du schon, Schmerz?
Was mich beglückte,
Sage, was wars doch, mein Herz?

Freude wie Kummer,
Fühl ich, zerrann,
Aber den Schlummer
Führten sie leise heran.

Und im Entschweben,
Immer empor,
Kommt mir das Leben
Ganz wie ein Schlummerlied vor.
```

```
Strophe 1:
Silbenzerlegung:
  (Fried|lich) (be|käm|pfen)
  (Nacht) (sich) (und) (Tag) .
  (Wie) (das) (zu) (däm|pfen) ,
  (Wie) (das) (zu) (lö|sen) (ver|mag) !
Metrik:
  Silben=5, Betonung="+--+-"
  Silben=4, Betonung="+--+"
  Silben=5, Betonung="+--+-"
  Silben=7, Betonung="+--+--+"
Reim:
   Endreim="abab" (Kreuzreim)
Klassfikation nach Frank:
  unbekannt
Klassifikation nach Schlawe:
  4.73

Strophe 2:
Silbenzerlegung:
  (Der) (mich) (be|drück|te) ,
  (Schläfst) (du) (schon) , (Schmerz) ?
  (Was) (mich) (beg|lück|te) ,
  (Sa|ge) , (was) (wars) (doch) , (mein) (Herz) ?
Metrik:
  Silben=5, Betonung="+--+-"
  Silben=4, Betonung="+--+"
  Silben=5, Betonung="+--+-"
  Silben=7, Betonung="+--+--+"
Reim:
   Endreim="abab" (Kreuzreim)
```

```
Analyse durchgeführt:  Silben Metrik Reim Klassifikation(Frank) Klassifikation(Schlawe)
 Metrik: dakt(5) jamb(14) troch(14) dakt2(17) anpäst(18)
```

Abbildung 1. metricalizer 0.8: Links wird der entsprechende Text hineinkopiert, auf der rechten Seite erscheint das Analyseergebnis

[1] www.lyrik-und-lied.de; www.freiburger-anthologie.de

[2] Ähnliche Projekte für das Deutsche sind „YASP" <http://wiki.cl.uni-heidelberg.de/YetAnotherSyntheticPoem> und das Projekt „ErMaStat": Dimpel, Friedrich Michael: Computergestützte textstatistische Untersuchungen an mittelhochdeutschen Texten. 2004.

[3] Was „richtig" in diesem Zusammenhang bedeutet, ist hier nachzulesen (Fußnote 41): <http://computerphilologie.tu-darmstadt.de/jg07/bobgehl.html#FN41>

[4] Enwickelt von Regine Müller: http://www.cross-plus-a.com/segmenti.htm

[5] Friedrich Hebbel: Abendlied . zitiert nach <http://www.lyrik-und-lied.de/ll.pl?kat=typ.show.poem&ds=2580&id=4082>

[6] Schlawe, Fritz: Die deutschen Strophenformen. Systematisch-chronologisches Register zur deutschen Lyrik 1600-1950. 1972. / Frank, Horst Joachim: Handbuch der deutschen Strophenformen. Tübingen/Basel. 1993.

[7] Eine Testversion des Programms ist nun online unter: <http://www.poetron-zone.de/metricalizer/generator.php>.

[8] „Regelmäßige" Gedichte sind nach unserem Verständnis Gedichte, die aus einer strophenweise in sich regelmäßigen Abfolge von betonten und unbetonten Silben bestehen, die mit dem gleichen Muster pro Vers beginnen müssen, aber unterschiedliche Endpunkte des Musters zulassen. Ein Muster ist die kleinste sich wiederholende Struktur von betonten und unbetonten Silben innerhalb eines Verses (im Prinzip identisch mit einem Versfuß).

[9] Eine detaillierte Beschreibung des Projektes findet sich nun online unter: <http://computerphilologie.tu-darmstadt.de/jg07/bobgehl.html>. Der gesamte Regelsatz ist noch nicht für eine Außendarstellung vorbereitet.

No Job for Techies: Collaborative Modeling as an intellectual activity of the analyst and scholar in the development of formal representations of scholarly materials.

John Bradley
King's College London
john.bradley@kcl.ac.uk

This paper deals presents some thinking about the nature of the workthat are involved in the development of digital resources for the humanities; and the participants in that work. It has become almost a truism among many in the digital humanities (DH) that digital resource building is, of necessity, a kind of collaborative activity. The scholar who is sponsoring the resource brings to the table the materials that s/he wishes to work with, and contributes the issues that these materials raise that are of scholarly importance. However, s/he is not going to be in the position to grapple his/herself with the myriad technical matters that arise from using the technology to represent them. Instead, as this truism goes, s/he will need technical *support* to help him/her implement his/her vision and present it. In this view of things, the partnership is based on the scholar being responsible for the content, and the technologist who is given the job of representing the vision of the scholar. This view of the resulting partnership is not equal in nature (indeed, the term 'techie', which I hear surprisingly often applied to the technical person in this kind of partnership, confirms this), and it is based on the scholar as *faculty* and the technician as *staff*. It is widespread and influences much of what has been written about the collaboration models that currently exist. See, for example Zorich 2007-8, or Michel et al. 2003, in which there is both frequent reference to "faculty" and "technician", and where the main interest in actual *intellectual* collaboration is not between the scholar and his technical support, but the different one which might be developed between academics from different disciplines (such as between History and a Fine Arts department, or, interestingly, between a humanities department and computer science).

The Centre for Computing in the Humanities at King's College London (CCH) operates differently. The first sign of this is that CCH is set up as a full *academic* department within the School of Humanities. Thus, it has a combined teaching and research mission not unlike other academic departments – although in the case of CCH its research is based around the Digital Humanities rather than any conventional humanities discipline. Indeed, CCH has been able to take creative advantage of the UK academic context which does not divide its staff quite so clearly into faculty and staff.

Resources developed at CCH are, of course, still meant to express the scholarly interpretive inputs of their scholarly partners, but the work to achieve this is carried out with extended intellectual collaboration between us and our discipline partners. Resources that emerge from this process are something like John Unsworth's *Thematic Research Collections* (Unsworth 2000) – objects defined by Carole L. Palmer as 'digital aggregations of primary sources and related materials that support research […]' (Palmer 2004, p. 348), even if some of them, for example the several prosopographies in which we have been involved, cannot be properly described as 'digital aggregations of primary sources'.

Another key difference is in how we name, and therefore consider, our technical staff. We view our key contributors (both in the context of XML and TEI, or in the context of database design) not at technicians, but as *analysts*, with significant intellectual input to the projects they work on. The work that emerges from the scholar/analyst partnership shares something with the kind of software design process described by the Scandinavian Design School (see Greenbaum and Kyng 1991 for an extensive introduction), who are interested in what they describe as 'the sociological and anthropological area of system design' (Sissen 1998 p. 11). One of the design strategies they have identified they call *Collaborative* (or *Participatory*) *Design*.

For the CCH Analyst the work centers on the central task of *modelling* the data (largely in the sense of McCarty's view of modelling (McCarty 2004)). In our resource-development projects we try to develop a digital representation of some significant part of the scholar's intellectual model by formalising it. By expressing some part of the interpretative model in sufficiently formal terms, the computer – and therefore the digital tools that we develop to present the materials to end users – can better exploit it to provide an enriched kind of interaction. This formal model is wrapped up tightly with the resource objects that the project delivers, and formally represents some significant aspects of the scholar's interpretation of his/her materials. The modelling task is not so much one of simply applying an existing model (such as, say, Dublin Core or even TEI or CIDOC, although these are, of course, often used as building blocks) to materials of interest to our scholar-partners, but to develop a model that specifically represents the interpretative framework that the scholars themselves are developing.

Furthermore, the resulting model does not emerge solely from the scholar speaking about what is to go into it and the analyst writing it down. Instead, the model develops as a bilateral collaborative activity. The scholar, of course, brings a deep knowledge of the subject, and in particular an understanding of where the thorny bits lie. The analyst, in turn, brings an understanding of formal methods of modelling and combines this with an experience of dealing with complex humanities material that has arisen from other projects in which s/he has participated. S/he also acts to clarify structures (naming, attributes and relationships) that are essential to the model. Out of this comes new understanding for *both* the scholar and analyst partners. As Finken (1998) says:

> Cooperative Design has […] a central and continual discourse about egalitarity [sic]. This presupposes that users and designers enter a work setting of mutual learning, where they are equal partners; the users are said to be skilled experts, the designers are technological experts. (p 6)

The amount of work, and exchanging of issues and ideas in both direction shows that the model is a joint intellectual product. Our partners sometimes tell us that the resulting model has ideas sufficiently intertwined from both the partners that it is not possible to separate them.

Since modelling is both an intellectual activity, and a collaborative one, at CCH we have found it important to recognise the work of the analyst as am intellectual activity that should usefully be considered as research or at least research-like. Therefore, as with scholarly research output, the analyst's work also benefits from the sharing of it with other professionals. As Finken says about developers:

> The practitioners are not unambiguous developers, designers and/or technological experts; they are scientific designers, which implies that they make a living by *doing research*. This means, that the scientific designers test different methods, techniques or theoretical hypothesis during a period of time, and then later write about their experiences. So besides doing system design, the designers also address a community of researchers, who also have interests in the scientific side of system design. (Finken 1998, p. 8)

How, then, could the academy foster an environment where such intellectual activity can occur? First, it should provide an environment that sustains these kinds of intellectual partnerships, and where a kind of peer relationship between the designer and the scholar is supported and acknowledged. Second, it should recognise that the work of the analyst contains elements of research in its own right, and that the analyst needs to be recognised as a researcher in broadly the same way as the academic is. In other words, the analyst may not be *faculty*, but they are not *support staff* either.

CCH's role as an academic department within Humanities, and the role of its staff analysts and developers as, at least in part, academic individuals has meant that CCH is particularly well placed to explore some of these issues. For example, all CCH staff are encouraged to do research or research-like work that can result in traditional outputs such as conference papers or articles. However, both CCH and the School of Humanities at KCL are still struggling with some the implications of this. What, exactly, is meant by *research* for the modelling analyst? How can research time be found for the analyst, and what are the appropriate research outputs for these individuals? How does this work fit into the teaching mandate for CCH?

My presentation will draw on examples from our experience to illustrate how the analyst role, even though technical in orientation, participates in the intellectual and academic development of digital resources, and how this technical work can be made more evidently into an intellectual one with the potential to appear as research. In this way, the analyst's contribution shares, in some ways at least, many of the aspects of the scholarly output of our projects' discipline-based partners.

References

Finken, Sisse (1998). *Truth Is a Thing of This World: A Foucaultian Analysis of the Discoursive Construction and Constitution of Cooperative Design*. Available online through http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.3190. Accessed 7 November 2008.

Greenbaum, Joan and Morten Kyng (eds) (1991). *Design at Work: Cooperative Design of Computer Systems*. Boca Raton, FL: CRC Press.

Palmer, Carole L. (2004). "Thematic Research Collections". In Schreibman, Susan, Ray Siemens and John Unsworth (eds), *A Companion to Digital Humanities*. Malden, MA: Blackwell Publishing. pp. 348-365. Also online at http://www.digitalhumanities.org/companion/ (chapter 24). Accessed 7 November 2008.

McCarty, Willard (2004). "Modeling: A Study in Words and Meanings". In Schreibman, Susan, Ray Siemens and John Unsworth (eds), *A Companion to Digital Humanities*. Malden, MA: Blackwell Publishing. pp. 254-270. Also online at http://www.digitalhumanities.org/companion/ (chapter 19). Accessed 7 November 2008.

Mitchell, William J., Alan S. Inouye and Marjory S. Blumenthal (eds) (2003). *Beyond Productivity: Information Technology, Innovation, and Creativity*. A report for the National Research Council of the National Academics. Washington D.C.: The National Academies Press. Pp. 251.

Unsworth, John. (2000). *Thematic Research Collections*. Paper presented at the Modern Language Association Annual Conference, December 28, Washington, DC. Available online at http://iath.virginia.edu/~jmu2m/MLA.00/. Accessed 7 November 2008.

Zorich, Diane M. (2007, 2008). *A Survey of Digital Humanities Centers in the United States*. Council on Library and Information Resources. Online at http://www.uvasci.org/wp-content/uploads/2008/06/dhc-survey-final-rept-2008_05_22-for-distribution.pdf. Accessed 7 November 2008.

On-site Scanning of 3D Manuscripts

Timothy H. Brom
University of Kentucky
thb@netlab.uky.edu

James Griffioen
University of Kentucky
griff@netlab.uky.edu

W. Brent Seales
University of Kentucky
seales@dcs.uky.edu

Introduction

Library collections all over the world contain manuscripts, scrolls, and other documents that have never been read because they cannot be opened in a way that exposes the text without causing physical damage. To address this problem, the EDUCE project at the University of Kentucky [2, 3] is experimenting with the use of Computed Tomography (CT) scanners to read fragile scrolls that cannot be unrolled and fragile manuscripts whose content is inaccessible because the pages cannot be separated.

The basic idea at the heart of the EDUCE project is to use a high-resolution CT scanner to "see inside" the scroll or manuscript. Given the raw CT scan data, computers can compute a 3-dimensional image of the artifact, called a voxel set. Computation of the voxel set is commonly referred to as reconstruction. Computers can then be used to look for structures in the image (e.g., to find where the rolls of the scroll are in the voxel set). Given the structure of the document, the computer can focus its search for traces of ink, discerning ink from parchment. Having identified the ink, algorithms can be applied that "virtually unroll" the document to reveal the hidden text.

The Reconstruction Problem

Although the EDUCE approach involves several computational steps, by far the most time consuming step is the reconstruction step. Reconstruction is compute intensive and can run for hours or days depending on the size of the data and the speed of the computer. As such, reconstruction times have the biggest effect on the rate at which artifacts can be scanned and ingested into a digital collection. Moreover, it often take several "trial" scans to find the best settings for the scanner, which means that reconstruction must occur several times (as part of these "trials") before the "real" scan can proceed. Consequently, reducing reconstruction times is critical to the viability of the EDUCE approach. Our initial estimates to scan relatively small objects were on the order of weeks. This is a major problem, particularly when the artifact can only be out of storage for a short time.

Current solutions to the reconstruction problem include (1) low-resolution reconstructions, (2) partial reconstructions, and (3) full reconstructions performed on a high-performance cluster or supercomputer. Unfortunately, low-resolution scans often fail to identify the tiny ink particles that make up the text. The second alternative, high-resolution reconstruction of part of the artifact, can be done faster than reconstructing the entire artifact, but it can miss the region containing the text, or can produce scan settings that are not useful across the entire artifact.

By using large compute clusters, one can harness enough computer power to perform a high-resolution reconstruction of the artifact in a relatively short amount of time. However, it means the (fragile) artifact must be transported to the location where the large cluster (and its associated scanner) reside. Unfortunately, this is often impossible because the artifact is not in a condition to travel. The alternative is for the cluster (and its scanner) to travel to the location where the artifact is housed. Historically, however, compute clusters and scanners have not been designed to be mobile. Consequently, the EDUCE project has been exploring ways to make both the scanner and the compute cluster portable.

Fortunately, recent advances in scanner technology have produced portable CT scanners whose scanning volume can hold reasonably large objects (say a book or scroll). Moreover, these scanners now are capable of the high-resolution scanning that is needed to discern the substrate that the layers of the scroll are composed of (papyrus, vellum, paper, etc) which are very thin, especially in damaged documents where layers are fused together. One such example are the portable scanners from Skyscan [1] which are both small and light enough to be portable.

However, the ability to perform computationally intense reconstruction processing on the resulting data without the aid of a large compute cluster remains a challenge. In the remainder of this paper, we focus on the problem of creating a small-scale, portable computation system that is capable of performing fast reconstruction.

Portable Reconstruction

One way to address the problem is to utilize remote computational power by copying the scanned data over the

Internet to a remote computer for reconstruction. Given the massive amounts of data produced by a single scan–on the order of tens of gigbytes–one would need a very high-speed network connection for this approach to be viable. Moreover, "trial" scans only increase the number of times that the data must be moved to the remote cluster for processing. The network also becomes a single point of failure, causing work to halt if there is a network outage. For these reasons–and because we did not want to rely on the institutions we were visiting to have a reliable high-speed internet connection–we turned our attention to developing a small, portable compute resource that we could take with us on-site.

Our initial testing of the CT technique were done with a compute cluster of 64 machines. While this worked quite well for the immediate task at hand (testing the viability of the general concept), a cluster of 64 machines is decidedly not portable. Since portability was a necessary goal for this project, given that it was not possible to bring the artifacts to be scanned to us, a new cluster design was needed.

The recent advent of multi-core CPUs has provided a significant boost in the amount of computing power available in a single machine. Consequently, we decided to build our portable cluster out of four multi-core computers, each with dual quad-core processors, yielding up to 32 processing cores to apply to the reconstruction task. We then spent significant time modifying and optimizing the code to run on a multi-core architecture where parallel memory and disk I/O turned out to be bottlenecks. Although one could argue that four physical PCs is not particularly portable, we were able to ship them to the scan site and set them up quickly with relative ease. Despite the greatly reduced size (as compared to the original 64 node cluster), the new cluster ordered excellent performance; performance on par with the orginal cluster.

We were still interested in trying to make our setup even more portable. We decided to explore the possibility of utilizing a Graphics Processing Unit (GPU) to increase the speed of the reconstruction computations. NVidia recently released a toolkit known as the *Compute Unified Device Architecture* (CUDA) [4] which allows code written in C or C++ to be ported to run on a GPU. Since the problem of CT reconstruction is highly data-parallel, and the GPU is optimized for data-parallel computation, we hypothesized that the GPU could provide speedups over the same program running on a traditional CPU.

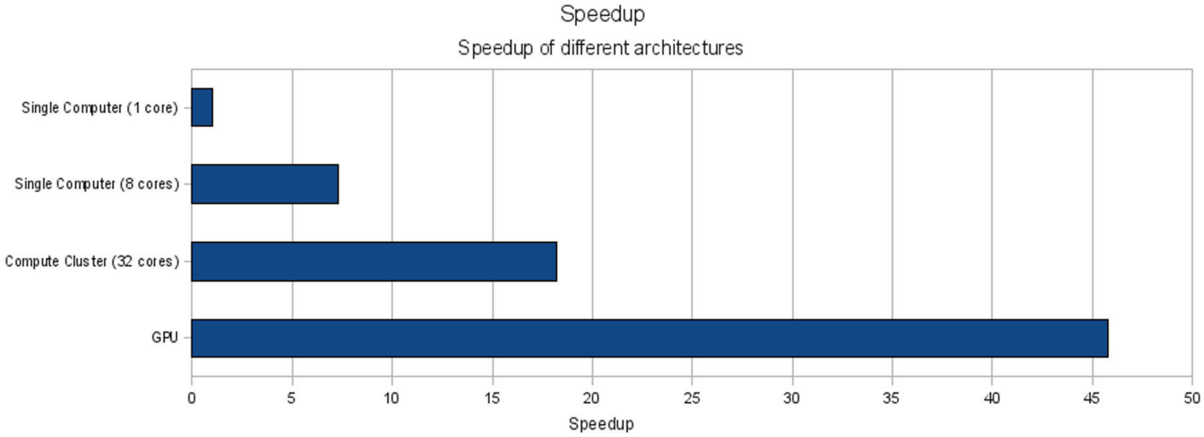


Figure 1. Comparison of the speedups obtained by different compute architectures for the CT reconstruction problem

Porting the code to run on the GPU was non-trivial, largely due to the need to carefully place (and move data around) in memory. However, the performance speedups were quite impressive. A single GPU outperformed our entire cluster of computers (32 processing cores) by a factor of more than 2 to 1. In other words, a single laptop computer with a powerful CUDA-capable graphics card can be used to run our reconstruction code fast enough for most reconstruction jobs, and a single computer that contained multiple GPUs could replace large clusters of machines.

Another advantage of using the GPU for computation is reduced cost of the computer equipment. A single machine with a high-end graphics card or a similarly equipped laptop are both significantly cheaper than a cluster of computers. A single machine is also simpler from a technical standpoint, a cluster brings with it an additional communication network and software layers which are unnecessary for a single machine.

Being able to run a high-resolution CT reconstruction on a laptop in a reasonable amount of time eliminates the problem of portability for the computational requirements of CT scanning, and even transporting a single computer that contained multiple graphics cards for larger jobs is often feasible. This, coupled with the commercial availability of CT scanners that are small enough to be portable, makes this technology quite feasible for digitally exploring artifacts.

Conclusion

In this paper, we studied the problem of reconstructing damaged manuscripts and described the reconstruction problem that must be solved in order to achieve on-site portable scanning of an artifact. We briefly described two ways in which the necessary computational power can be achieved using recent advances in multi-core architectures and graphics processing units. The performance of the different solutions considered were reported, and indicate that a single GPU can be used to order performance that was previously only available on large clusters of PCs.

This work supported in part by the National Science Foundation under Grant No. IIS-0535003.

References

[1] Skyscan. http://www.skyscan.be/home.htm.

[2] Educe: Enhanced Digital Unwrapping for Conservation and Exploration, 2006-2009. http://www.stoa.org/educe.

[3] Alicia P. Gregory. Digital Exploration: Unwrapping the Secrets of Damaged manuscripts, 2004. http://www.research.uky.edu/odyssey/fall04/seales.html.

[4] NVidia Corporation, 2701 San Tomas Expressway, Santa Clara, CA. *NVIDIA CUDA Compute Unified Device Architecture Programming Guide*, 2.0 edition.

Modeling the Lexicon with Ontologies

Kip Canfield
University of Maryland, UMBC
canfield@umbc.edu

Introduction

The use of markup languages (typically XML) for modeling the natural language lexicon is currently widespread. It allows a standard representation format that is friendly to the web, interoperable, and leverages horizontal standards in that all general XML tools can be used (Simons 2004). This paper makes an argument for elaborating this practice to use ontologies for modeling lexicons. These would be commonly serialized in OWL/XML, but could use the standard XML transformation capability to use other serializations without any problem. The Lexical Markup Framework (LMF) is a popular ISO standard candidate for lexicons and shows the benefits described above. Since LMF is XML and transformations can allow one schema to be changed into another, using ontologies in this context has been suggested such as using an LMF to OWL transformation (Francopoulo 2007) to incorporate ontologies into a service-oriented language infrastructure (Hayashi 2008). This paper proposes using ontologies in a much more local way to support development and testing of natural language lexicons, but at the same time supporting the goals of interoperability and standards use. The argument for this novel use of the semantic web technology is advanced using a scenario from the development of a computational lexicon of the Navajo language of North America.

Methods

In the context of a long running project to create a computational resource for the Navajo lexicon that can be used for reading and annotating a corpus of Navajo texts, an XML-based model of the Navajo lexicon was developed. This lexicon has a native format of an XML schema, but leverages transformations to customize the format for any particular use. For example, the web based applications that display texts from the corpus and allow access to the lexicon, use a JSON transform since that is most convenient for web applications. Similarly, the lexicon can be transformed to OWL for creating an ontology. The ontology can be used in much the same way as LMF documents or databases have been used (Beck 2007), but with additional useful properties.

The class structure from the original lexicon model is easily transformed to OWL classes. Navajo verbs show a template structure with complex morphology that can be seen as a slot/filler structure were only certain kinds of affixes can be put into any slot (Young 1992; Faltz 1998). So the Navajo verb class has an ordered sequence of properties that correspond to the template (affix order=outer, distributed plural, inner, object, subject, classifier, stem). Some of the slot content can be predicted from context and some must be defined in the lexicon. For an OWL ontology, this would be the information in the individual. For example, the individual for the class verb "cut it out" would be:

(1) individual V2 is Verb and OuterPx=ha, Transitive=1, Conjugation=S, Classifier=ł, Sid=2, Gloss="cut it out";

These examples use the non-XML serialization from the Bossam reasoner (Jang 2004) for ease of human reading. Two other classes are defined for the stem set and the subject prefixes. A stem set is a set of bound stems that change shape depending on the verb mode and aspect. A sample individual stem from the class stem set (StemSet) is:

(2) individual S3 is StemSet and Sid=2, Root=gizh,
Mode=I, Aspect=cont, Stem=géésh;

The stem in (2) is only for verbs with imperfective mode and continuative aspect and there would be 5 stems in the set. The subject prefixes vary widely depending on the conjugation type and other parameters. A sample individual from the class subject prefix (SubPxSet) is:

(3) individual Sj2 is SubPxSet and Mode=I,
Conjugation=S, Pnum=3, SubjPx=Ø;

This subject prefix is for the 'Simple' conjugation and the third person. The main point being that many of the parts of an underlying form for the verb (1) are predictable and so do not appear in the lexical entry. In order to supply these predictable parts, the semantic web extension for ontologies - the Semantic Web Rule Language (SWRL) is used. This allows rules to apply reasoning to the individuals in the ontology. For example, since the distributive plural is only for 3 or more, a simple rule using unification can add that prefix of 'da.' (Pnum=8 is the person number for the 8th member of a listing of the conjugation. Navajo has both a 3rd and 4th person, so the 9th would be the first member after the dual plurals.)

(4) rule Rule1 is if Verb(?v) and Pnum(?x) and [?x>8]
then DistPlPx(?v, da);

This also uses the Bossam rule serialization for ease of reading, but it can also read the XML serialization of SWRL. Similarly all the predictable parts of the underlying verb form are generated. For example, the verb for "cut it out" in the third person distributive plural is generated by the rule base:

(5) OuterPx= nav:ha, DistPlPx= nav:da, ObjPx= nav:y,
InnerPx= nav:Ø, SubjPx= nav:Ø, Classifier= nav:ł,
Stem= nav:géésh

The output in (5) shows that each slot in the template has been correctly filled using the information in the lexical entry and using rules (nav: is the namespace). The underlying form for that verb is ha-da-y-Ø-Ø-ł-géésh. This can in turn be transformed to the surface form using a set of morphophonemic rules that have been previously developed for this project. It uses the finite state morphology tool xfst (Beesley 2003) and produces hadeiłgéésh which is the distributive third person of the surface form. This is a command line tool and so the above can be done with a single unix pipeline. This is an effective tool for checking lexical entries. For example, a complete paradigm can be cosnstructed using the lexical ontology rules to produce the underlying form (left-side of (6)) and xfst to produce the surface form (right-side of (6)) :

(6) Complete imperfective paradigm for: individual P5
is Verb and OuterPx=Ø, Transitive=0, Conjugation=S,
Classifier=Ø, Sid=5, Gloss="cry";

1sing: Ø-Ø-Ø-Ø-sh-Ø-cha -> yishcha
2sing: Ø-Ø-Ø-Ø-ni-Ø-cha -> nicha
3sing: Ø-Ø-Ø-Ø-Ø-Ø-cha -> yicha
4sing: Ø-Ø-Ø-Ø-j-Ø-cha -> jicha
1dual: Ø-Ø-Ø-Ø-iid-Ø-cha -> yiicha
2dual: Ø-Ø-Ø-Ø-oh-Ø-cha -> wohcha
1distpl: Ø-Ø-Ø-da-iid-Ø-cha -> deiicha
2distpl:Ø-Ø-Ø -da-oh-Ø-cha -> daohcha
3distpl: Ø-Ø-Ø-da-Ø-cha -> daacha
4distpl: Ø-Ø-Ø-da-j-Ø-cha -> dajicha

Analysis and Conclusions

Using ontologies for modeling the lexicon has benefits at both the local, lexicographer level and the global level for searching and interoperation. The approach has particular benefit for languages with complex morphology. As seen in this example from the Navajo language, the process for setting up an underlying form has many parts that can be modeled in the lexicon using the Semantic Web technology. Typically, lexical entries would be isolated for grammatical exposition, but with this approach, the lexicon becomes an active resource that incorporates those grammatical facts. Furthermore, the technology is horizontal in nature, meaning that the tools used are not particular to this application or even this class of applications. That makes it is easier for everyone to understand it and share it based on familiarity with the general technology of the Semantic Web. This helps to avoid the short ‘half-life’ common to computational lexicography projects (Maxwell 2008).

More globally, this approach allows the lexical model to participate in the more common applications of the Semantic Web such as annotation, search, and linking resources on the Internet. The resource can also participate in global semantic relations as with WordNet. For the Navajo lexicon described here, connecting the model to other ontologies was made simple. For example, each class is connected to the GOLD general OWL ontology for linguistic concepts (Farrar 2003) dynamically over the Internet using the built-in OWL property owl:sameAs which links an individual to an individual. Since much of the terminology in any particular language’s linguistic literature can be divergent, this helps interoperation (Chiarcos 2008).

References

Beck, H. (2007). *Contextual Archiving with Linguistic Analysis: An Ontology-Based Approach to Developing a Linguistic Database*. Workshop Proceedings of “*Toward the Interoperability of Language Resources*”, July 13-15 at Stanford University in conjunction with the 2007 LSA Summer Institute, http://linguistlist.org/tilr/proceedings2.cfm, (accessed 11/11/2008).

Beesley, K. and Karttunen, L. (2003). *Finite State Morphology*. Stanford: CSLI Publications.

Chiarcos, C. (2008). An ontology of linguistic annotations. *LDV-Forum*, 23 (1) , 1-16, http://www.ldv-forum.org/?language=en, (accessed 11/11/2008).

Faltz, L. (1998). *The Navajo Verb*. Albuquerque: University of New Mexico Press.

Farrar, S. and Langendoen, T. (2003). A Linguistic Ontology for the Semantic Web. *GLOT International* 7(3), 97-100, http://www.linguistlist.org/emeld/documents/GLOT-LinguisticOntology.pdf, (accessed 11/11/2008).

Francopoulo, G. (2007). *Strategy for an OWL specification of LMF*. http://www.lexicalmarkupframework.org/ (accessed 11/11/2008).

Hayashi, Y., Declerck, T., Buitelaar, P. and Monachini, M. (2008). *Ontologies for a Global Language Infrastructure*. *The First International Conference on Global Interoperability for Language Resources* (ICGL2008), http://langrid.nict.go.jp/en/publication.html, (accessed 11/11/2008).

Jang, M. and Sohn, J. (2004). Bossam: An Extended Rule Engine for OWL Inferencing. Antoniou, G. and Boley, H.(Eds.): *RuleML 2004, LNCS 3323*, 128–138, http://bossam.wordpress.com/documentation/, (accessed 11/11/2008).

Simons, G., Lewis, W., Farrar, S., Langendoen, D., Fitzsimons, B. & Gonzalez, H. (2004). *The Semantics of Markup: Mapping Legacy Markup Schemas to a Common Semantics*. Proceedings of the XMLNLP Workshop, Association for Computational Linguistics, Barcelona, Spain, http://emeld.org/documents/SOMFinal1col.pdf, (accessed 11/11/2008).

Maxwell, M. and David, A. (2008). *Interoperable Grammars*. The First International Conference on Global Interoperability for Language Resources, Hong Kong, January, http://icgl.ctl.cityu.edu.hk/, (accessed 11/11/2008).

Young, R. and Morgan, W. (1992). *Analytical Lexicon of Navajo*. Albuquerque: University of New Mexico Press.

Lost in Transcription: Types, Tokens, and Modality in Document Representation

Paul Caton
National University of Ireland
pncaton@gmail.com

It is a commonplace that representation – no matter how detailed – loses something from the original; thus adequacy in representation is always contingent.[1] A recently proposed formal model of transcription (Huitfeldt and Sperberg-McQueen, 2008) describes the criterion that must be met if we are to say one document is a transcription of another. Starting from that model this paper first identifies a particular source of information loss between exemplar and transcription, then generalizes from that to a class of losses and suggests what the model should include to account for that class. Finally, the paper shows how aspects of the model might be realized in markup.

In the model an exemplar document E is a physical object on which is written a sequence of tokens, and there exists a reading of that sequence in which each token instantiates a type in a one-to-one relationship. (The model is agnostic as to the granularity of tokens.) A second document T is similar to E if and only if there is a reading of the token sequence in T such that for each token in T:

1. we know which token in E it corresponds to
2. the order of tokens in T matches the order of tokens in E, and
3. the type instantiated by the token matches the type instantiated by the corresponding token in E

When the type sequences of T and E match, the documents are t_similar, which is the model's formal requirement for T being a transcription of E. Note that the model includes no formal concept of information loss between E -> T other than that implied by a failure of any reading of T to generate a type sequence that matches the E type sequence. Figure 1 shows the basic model of transcription and how it is repeatable: if E and T_1 are t_similar, a transcript T_2 can be made of T_1 such that E and T_2 are t_similar.

The model defines a document as "an individual object containing marks" (297). On application of a reading some, none, or all of the marks may be identified as tokens insofar as they instantiate types. It is not unusual for many acts of transcription to begin with considerable uncertainty as to whether the marks in the document are tokens at all, but as our interest is in text encoding we shall assume an E where a competent reading not only identifies some or all of the marks as tokens but also recognizes that the token sequence forms a normative text: that is, a text that conforms to the morpho-syntactic and orthographic rules of the language from whose writing system the types are being instantiated. Note that we cannot assume meaningfulness in the token sequence: a sentence such as Chomsky's famous "Green ideas sleep furiously" still counts as a text.

The model *assumes* E from the start, but for our purposes we need to make the genesis of E explicit (even if only - because necessarily - in an imaginary way). Figure 2 shows the model extended backwards temporally to include a moment of instantiation that produces the E token sequence. We need not go into the question of whether there can actually be an uninstantiated type sequence,[2] but it is important to include the process of instantiation of the e_tokens – and what may happen in it - because there is no necessary identity *at the token level* between an e_token and its corresponding t_token. If we know the e_type sequence and we wish to create a t_similar document, we can only establish t_similarity by instantiating each e_type as a t_token in a manner that preserves the original e_type::e_token relation (subject, of course, to a suitable reading). In other words, t_similarity does not *depend* on T having an identical token sequence to E.

As noted above, the model deliberately remains agnostic as to the level at which tokens are distinguished, but we will follow Huitfeldt and Sperberg-McQueen and start by considering tokens at the level of the smallest individual units in a writing system., which for convenience we will refer to as the grapheme level.[3] Assume a document E that contains the text "How could Henry be *here*, when he is supposed to be at his house?" The "H" tokens in "How" and "Henry" are visibly different from the "h" tokens of "when", "he", "his" and "house". We assume that in E the latter four have the kind of accidental differences common to cursive handwriting or mechanical printing, but they are obviously intended to be seen as identical tokens and thus instantiations of the same type. The "H" is specifically not meant to be *seen* as identical to the "h": the deliberate difference is called for by modern English orthography. The choice of the token "H" is rule-governed, just as is the choice of the "?" to punctuate the end of a question. Majuscule and miniscule 'h' may be different glyphs, but they are allographs of the same grapheme and thus are tokens that instantiate the same type.[4]

To identify tokens at a higher level, we shall use the notion of the frame where meaningful units of tokens (either single or in groups) are made distinct either by a framing mark or the systematic absence of a mark (in modern English orthography, the whitespace).[5] Consider the frame-level token “*here*”. At the grapheme level, the token “*h*”, though not the same glyph as either the “h” of “when”, “he”, “his” and “house” or the “H” of “How” and “Henry”, is clearly still an allograph of the grapheme they instantiate and therefore instantiates the same type as they do. At the frame level, however, we see a difference that is not the same sort as that between “h” and “H”. In the process of instantiation a document creator has made a particular choice about the form of all the glyphs in “*here*” that is *not* orthographically rule-bound like the choice of “H” in “How”. We understand that the use of a different typeface at this point is not random: it is deliberate and has communicative intent, even if we have to make a more or less informed guess as to its precise significance.

Does “*here*” instantiate a different type than “here” (or “Here”, or “here”, etc.)? The frame-level token is a rule-governed grouping of grapheme-level tokens, and thus the frame-level type is similarly a rule-governed grouping of grapheme-level types. If “*h*” instantiates the same type as “h”, then “*here*” instantiates the same type as “here” and according to the model the following would count as a transcription of the text in E: “How could Henry be here, when he is supposed to be at his house?” Yet it is clear that in this transcription we have lost some information that was deliberately put into the token sequence of E. We must either account for this information in the notion of types themselves, or adjust the model to account for it somewhere else.

For reasons that have already been suggested, I believe we should resist locating this information at the type level. The notion that different token forms can map to one type is a strength of the model, and accords with our informal sense of what constitutes a transcription (as opposed to, say, a facsimile). Types *are* abstractions, but ones that derive precisely from the existence of variant forms. Any conception we have of the English alphabetic letter type ‘h’ has been shaped by the millions of ‘h’ tokens we have encountered. Similarly for the English morphological unit type ‘here’. It would seem odd to say that if we encounter the token “*here*” where in context we would expect “here”, then “*here*” must be instantiating a different type than “here”. If that were so, then what about “here”, or “**here**”: would they be two more different types?

That question points to what I consider the correct way forward. We would not consider it unusual if in our E text we saw “here”, or “**here**” instead of “*here*”; we would, I suggest, interpret the communicative intent the same way in all three cases. Supposing that the use of italics in “*here*” signifies emphasis, then we have a textual effect equivalent to a paralinguistic phenomenon. Just as emphasis in speech comes in delivery of the word, and is not part of the morphological unit *per se*, so the use of italics (or underlining, or bolding, or majuscules) are part of the ‘delivery’ – the instantiation – of the type.

As experienced readers we recognize a set of cases where tokens display something that is in excess of - or deviant from - the norm: that is non-rule-governed (though it may be conventional within a community of practice) and external to the type sequence. Let us call this thing *modality*, and let us distinguish three main types. The first, that we have already mentioned, is *presentational modality*. The variant token forms “here”, “**here**” and “*here*” all display presentational modality. The second kind is *accidental modality* and this, too, occurs in the process of instantiation. Examples of accidental modality would be turned letters, incorrect letters,, misaligned letters, broken typefaces, words out of sequence, etc. The third kind is *temporal modality*, and unlike the other two this occurs after instantiation of E but before the reading that generates the type sequence from which the token sequence in a transcription T will be instantiated. As the name suggests, this modality involves the effects of time on the token sequence in E, and includes staining, foxing, fading, darkening, blurring, etc. Figure 3 shows the model augmented with the three types of modality.

It should be clear that adding modality does not change the underlying model. One can choose to ignore the modal information and strive only for t_similarity. The idea that the modal information is less important and therefore ‘loseable’ accords with many peoples’ conceptions of transcription (and also, to some extent, of editorial practice and text encoding practice). We began by noting, however, that representational adequacy is contingent, and so for the creators and users of a scholarly digital edition t_similarity might be inadequate. Granting that it would be impossible to formally specify the difference between *t_similar* and *identical*, it still helps if the model includes the fact of - and something of the nature of – that difference. This is especially true when in many cases the reading has to interpret the modality to produce the type sequence.

In digital humanities practice, modality lost in transcription can be supplied by various means including images, prose descriptions, and the formatting facilities of text editing software. Encoding schemes such as the Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange (TEI Consortium, 2008) offer more abstract and more computationally exploitable ways of conveying modal information, but only as part of a whole whose scope is much wider than the particular concerns of the model. We could imagine, however, an encoding which in the model's terms we would locate between the reading and the T token sequence. This would be a highly abstract, contentless encoding that would, in effect, represent a *latent* token sequence and would record both type and modality (Figure 4).[6] In practice this encoding would mediate the final instantiation of types-into-tokens, giving the 'transcriber' in charge a high degree of control over how much information was preserved in the final token sequence.[7]

Notes

[1] Extensively commented upon; the following quote from Stevens and Burg is representative: "[t]ranscription is akin to translation, for no editor can take a document and convert it into another form without somehow changing it." (1997, 21). See also Shillingsburg (2008, passim) for an evocation of what is lost.

[2] Committed nominalists can simply treat the type sequence part of the model as a useful fiction.

[3] Certainly at this level the token::type relation closely resembles that of glyph::grapheme and it is convenient to use these familiar concepts. However, even in our imagination the abstract thing can only "show" itself according to tokens we have encountered, and thus it is hard to abstract graphemes as we should (see also following note). It may be that we need a more symbolic solution to the problem of specifying types.

[4] The model only works if we think of types as abstractly as possible, so while the alphabetic letters, punctuation marks, etc. available to English orthography have been historically determined and mutable over time, we have to think of them as members of a set, not as particular forms. There are 26 members of the subset we call the Latin alphabet, and I locate graphemes at this level, where the majuscule / miniscule distinction does not yet exist.

[5] I have taken this from DeFrancis (1989: 54), who is following Wang (1981: 226-228). DeFrancis contrasts English, where frames usually have more than one grapheme, with Chinese, where "frames invariably contain only one grapheme".

[6] The encoding shown in Figure 4 is simply a sketch of what transcription-oriented markup might look like; it is not meant to be taken as representing a fully worked out scheme.

[7] Of course the "final token sequence" can still be a mix of encoding and PCDATA.

References

DeFrancis, J. (1989) *Visible Speech: The Diverse Oneness of Writing Systems*. Honolulu: University of Hawaii Press.

Huitfeldt, C. and Sperberg-McQueen, C. M. (2008) What is Transcription? *Literary and Linguistic Computing*, 23: 295-310.

Shillingsburg, P. (1999) "Negotiating Conflicting Aims in Scholarly Editing: The Problem of Editorial Intentions." In Jansohn, C. ed. *Problems of Editing*. Pp. 1-8. Tübingen: Niemeyer, 1999.

Stevens, M. E. and Burg, S. B. (1997) *Editing Historical Documents*. A Handbook of Practice. American Association for State and Local History Book Series. Walnut Creek, CA.: AltaMira Press, 1997.

TEI Consortium, eds. (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. P5, Version 1.2.0. October 31st 2008. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (November 14th 2008).

Wang, W. S.-Y. (1981) "Language Structure and Optimal Orthography." In Ovid J. L. Tzeng and Harry Singer, eds., *Perception of Print. Reading Research in Experimental Psychology*. Pp. 223-236. Hillside, NJ.: Lawrence Erlbaum Associates.

Image as Markup: Adding Semantics to Manuscript Images

Hugh Cayless
New York University
hugh.cayless@nyu.edu

In their article "Images as the Text: Pictographs and Pictographic Logic," Jerome McGann and Johanna Drucker argue for the semantic significance of the image of text, using as an example the organization of the text of a poem by Byron. They conclude:

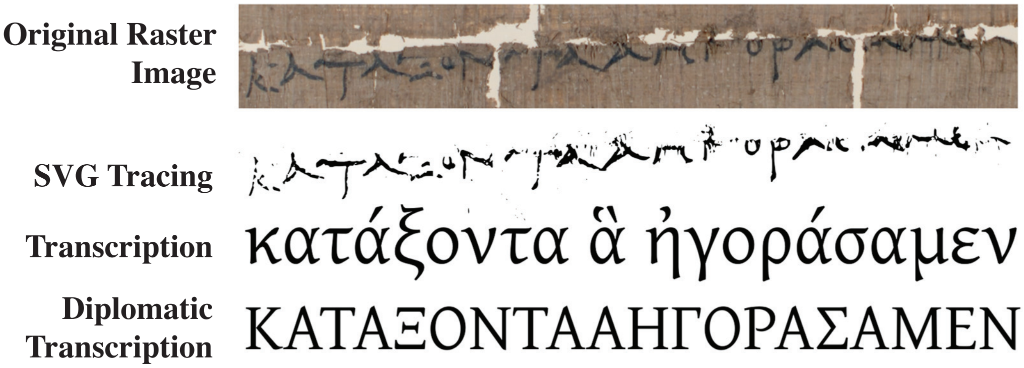
> ...a rhetoric of transparency makes it difficult to see beyond the moves within the text and the image to understand the metagraphic logic organizing them. In such a situation, the study of pictographs, which hover at the borderland of text and image, can be especially useful. They help us to see that at the next level of abstraction, of texts and images as information, similar logical mechanisms are at work. Each instance of text and image is an incarnation of such a metalogic, but it can be articulated according to its own rules and principles if it is rendered explicit. [McGann]

In the publication of manuscript transcriptions, two modes of presentation are typically recognized: the edition, in which an editor's supplements are folded into the text, and the diplomatic transcription, which attempts to faithfully reproduce the text on the original support. With the advent of markup systems like the Text Encoding Initiative (TEI), it is possible to produce both types of transcription from the same marked up text. Indeed, it is possible to go further, and analyze the text in ways that print transcriptions cannot, and to link transcriptions (and notes) to images in new ways.

As part of a series of experiments in text and image linking, beginning in the summer of 2008, the author has developed a method for generating a Scaleable Vector Graphics (SVG) representation of the text in an image of a manuscript. [CaylessSVG] This method employs open source tools to generate and present the results of the SVG tracing. Automated analysis of the SVG output of the process is capable (even using a naïve algorithm) of detecting lines in the source, and it is not hard to conceive of ways to detect words and other features in the image. The output of the tracing is in the form of SVG path elements, which employ a combination of cubic Bézier curves [Bézier] and lines to draw shapes. These can be grouped together (using svg:g elements) to, forexample, gather the components of a line of text as children of a single element. The SVG may include a copy of the original image as background, therefore superimposing the vector graphic tracings on top of the raster image. Since SVG is an XML application, this means that the features it represents are manipulable with Javascript, offering the possibility of highlighting features, panning and zooming, adding hyperlinks, etc.

In examining ways one might link between an XML transcription of the text and an XML overlay of the text, one quickly runs into problems involving overlapping hierarchies: paths may include multiple letters or words, for example, and there may be letters that correspond to multiple paths. The process of generating the SVG tracing involves the conversion of the image to a black and white (1-bit) bitmap, wherein each pixel is either black or white. This makes it possible for the software to reproduce the shapes in the original source in vector format, but it also involves a flattening of the text in the image into a single space. While it might have been clear that the stroke of one letter runs over the top of a second in a color image, that layering is lost in the SVG, and the two letters are a single shape in the output. This may, of course, involve the descender of a letter 'f' touching the line below, for example, making line detection more difficult.

The figure below (derived from a papyrus fragment[1]) highlights some of these issues. Notice that the initial kappa is represented by no less than eight paths, while part of the downward stroke of the final alpha in



κατάξοντα connects to the following word, ἅ.

These features of the traced text mean that marking a word in the image is inherently difficult. Possible solutions include modifying the SVG with an editor, so that the two letters no longer connect, or adding a new element, perhaps just a line added by an editor, to divide the two. The former solution involves doing a certain amount of violence to the image, however, since the two letters do in fact touch, while the latter introduces a new issue: the lack of semantics.

The semantics of SVG are almost purely geometric. It primarily encodes shapes, with additional support for links, text, embedded images, and animation. This means there is no inherent way to express the significance of a grouping or a feature in SVG. If we introduce the idea of a line that can break paths representing letters or words, then in order to be able to use this feature to, for example, point to the whole word κατάξοντα in the SVG, we would first need a process that could find the intersection points of the word-dividing line and the path representing the two alphas, and then split that path into two derivative paths, each of which would be associated with a different word. This method would avoid damage to the actual tracing while allowing the types of reference that are likely to be useful. The derivative paths could be placed in the same document and only activated as needed.

Again, however, there is a need for semantics in the SVG document. Not only might it be necessary to differentiate between lines dividing letters and lines dividing words or lines of text so that a processor knows how to deal with them and their outputs, but it must also be possible to distinguish between the derived paths and the originals, since there is no inherent difference between one svg:path element and another. They may render in the same coordinate space, for example, even though they are in different parts of the document.

Some possibilities for adding semantics to SVG include embedding metadata (perhaps using RDF) using the svg:metadata element, or developing a microformat, perhaps depending on the @class attribute, which is available on all displayable elements.[Schepers] These kinds of semantic "hooks" will be absolutely necessary if linking between the many possible structures in the transcription and the SVG are to be achieved. Some examples of the types of functionality that can be enabled by the technology described here are:

1. linking from notes, such as a description of letter forms, to actual examples inthe image.
2. The ability to link/zoom to any part of the text in the image from the transcription.
3. Image search with highlighted results.
4. Marking editorial emendations, such as expanded abbreviations and other types of editorial addition or deletion on the image itself.

Put another way, the semantics of the graphical representation of the text can be made explicit, by means of embedding and linking information into markup that itself overlays an image of the text, making for a very rich presentation and research tool.

Notes

[1] See http://papyri.info//navigator/full/apis_michigan_1769

References

Bézier curve - Wikipedia, the free encyclopedia. Retrieved November 14, 2008, from http:// en.wikipedia.org/wiki/B%C3%A9zier_curve. [Bézier]

Cayless, H. "Linking Page Images to Transcriptions with SVG." Presented at Balisage: The Markup Conference, 12 - 15 August 2008. In Proceedings of Balisage: The Markup Conference (2008). Retrieved November 14, 2008, from http://www.balisage.net/ Proceedings/html/2008/Cayless01/Balisage2008-Cayless01.html.

Drucker, J., & McGann, J. "Images as the Text: Pictographs and Pictorgraphic Logic." Retrieved October 28, 2008, from http://jefferson.village.virginia.edu/%7Ejjm2f/old/ pictograph.html. [McGann]

Schepers, D. A. "Reinventing Fire >> Blog Archive >> SVG Text, Semantics, and Accessibility." November 7th, 2006 at 5:24 am. Retrieved November 13, 2008, from http://schepers.cc/? p=11. [Schepers]

On Building a Full-Text Digital Library of Land Deeds of Taiwan

Jieh Hsiang
National Taiwan University
jhsiang@ntu.edu.tw

Szu-Pei Chen
National Taiwan University
gail@turing.csie.ntu.edu.tw

Hsieh Chang Tu
National Taiwan University
tu@turing.csie.ntu.edu.tw

In this paper we present a full-text digital library of Taiwanese land deeds. Land deeds were the only proof of land activities such as transaction of ownership and leasing in Taiwan before 1900. They form a major part of the primary documents at the grassroot level in pre-1900 Taiwan, and are extremely valuable for studying the evolution of the Taiwanese society.

Land deeds, on the other hand, are difficult to study because they are hand-written and hard to read. Furthermore, they are scattered in many different locations and, in some cases, in the hands of families and private collectors.

In order to make the land deeds more accessible to researchers and educators, the Council for Cultural Affairs of Taiwan embarked on a major effort to organize available land deeds and typed them as machine readable full-text. Based on this collection and collections from other sources, the National Taiwan University built a full-text digital library of Taiwanese land deeds. The current size of the collection is over 23,000 which, according to one estimation, cover about 50% of all existing land deeds.[1] The collection will be expanded to around 30,000 by the end of the year.

Our digital library is built with the goal of providing an electronic research environment for historians to conduct research using land deeds. Thus in addition to providing full-text search and retrieval, we developed a concept of regarding the query return as a *sub-collection* and built tools to help the user find meaning and relationships at the collection level. Post-processing presentation, term frequency analysis and co-occurrence, and relation graphs are some of the tools described in this paper. We believe that our digital library will bring Taiwanese historical research using land deeds to a different horizon.

Land Deeds of Taiwan

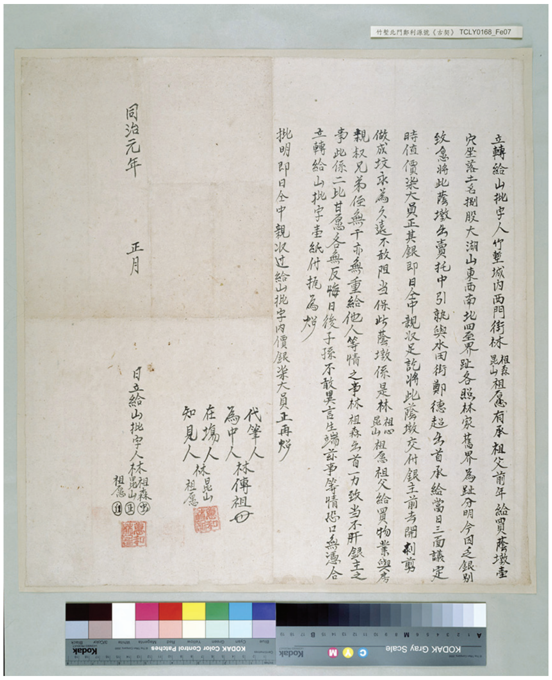
Before Taiwan was ceded to Japan by the Chinese Qing Dynasty after the Sino-Japanese War of 1895, land deeds were the only proof of ownership and transaction of land in Taiwan. Land deeds were, thus, a centrally important primary source for studying the 300 years written history of pre-1900 Taiwanese society (Wu, Ang, Lee, Lin, 2004). Even after the modernization of land administration by the Japanese, many families still kept the old land deeds either as part of the family heritage or due to the mistrust of the government. Many, however, were destroyed or discarded since they lost their original significance. During the early stage of the Japanese occupation, the government conducted research on the old administrative systems and customs and produced three series of books totalling 40 volumes, many of which contained transcriptions of land deeds as examples. During the reform of the land administration, the Japanese government sent surveyors to systematically transcribe land deeds so that they can be convert into modern land administrative records. In the latter endeavor, about 16,000 were collected. Their transcribed versions (copied verbatim by hand) are scattered in the 13,855 volumes of the Archives of the Japanese Taiwan Governor-Generals (Wang 1993). After Taiwan was returned to Chinese rule in 1945 after the 2nd World War, some research institutes and researchers recognized the importance of the land deeds and made efforts to collect them. The most notable, and largest scale such effort was conducted between 1976 and 1983 by a team lead by the historian Wang Shih-Ching who, commissioned by the Committee for Taiwan Historical Studies, Association for Asian Studies, U.S.A., collected about 5,600 land deeds and published a six volume catalog *Taiwanese Historical Documents in Private Holdings* (Wang, 1977). The photocopies of the land deeds were bound into more than 100 volumes. Other notable collections were kept at the National Taiwan University, the Institute of Taiwanese History of the Academia Sinica, the National Taiwan Library, and by various private collectors. Scores of books containing the images of some land deeds have appeared in the past ten years. Wang estimated (1993, pp. 71) that there are 20,000 land deeds in the hands of private collectors, libraries, museums and research institutes that were not included in the official collections. That makes the total number of such land deeds about 40,000. Our experience in the past ten years of digitization tells us there should be more, although we cannot give a reasonable estimation.

What is in a Land Deed?

Land deeds of Taiwan are contracts about various actions involving lands, such as the commission by the government to cultivate previously un-owned land, the division of family properties, the transaction of ownership, the rental of farming right, the pawning of land, etc. They were hand written and were usually prepared by a scrivener. A land deed usually consists of the following elements:

- The type of the land deed: selling, renting, pawning, etc.
- the "seller" (or owner) of the land,
- the "buyer" (or lender) of the land,
- the location of the land and its boundaries (usually marked on all four directions using neighboring landmarks such as river, road, building, pond, or even trees),
- the cost: money, maybe accompanied by other properties such as houses, cows, storage sheds, and farming tools,
- the names of witnesses and the scrivener, and
- the date.

The following is an example of a typical land deed.



竹塹北門鄭利源號《古契》 TCLY0168_Fe07

立轉給山批字人竹塹城內西門街林祖森崑山祖愿有承祖父前年給買蔭墩壹
穴坐落土名搁股大湖山東西南北四至界趾各照林家舊界為趾分明今因乏銀別
致急將此蔭墩出賣托中引就與水田街鄭德超出首承給當日三面議定
時值價柒大員正其銀即日仝中親收足訖將此蔭墩交付銀主前去開刹剪
做成坟永為久遠不敢阻当保此蔭墩係是林祖心崑山祖愿祖父給買物業與房
親叔兄弟侄無干亦無重給他人等情之事林祖森出首一力致当不肝銀主之
事此係二比甘愿各無反悔日後子孫不敢異言生端滋事等情恐口無憑合
立轉給山批字壹紙付執為炤
批明即日仝中親收过給山批字內價銀柒大員正再炤

代筆人林傳祖
為中人
在塲人林崑山祖愿
知見人

同治元年　正月　日立給山批字人林祖森崑山祖愿

Fig. 1 A typical land deed of Taiwan

What Kind of Research can be done with Land Deeds

While each land deed may have significance only to its owner, the collection as a whole provides a fascinating glimpse into the pre-1900 Taiwanese grassroots society. Through these land deeds, one can study the development of a region, or the overall land management, society, economy, and law of pre-1900 Taiwan. Furthermore, since many of the deeds were contracts between indigenous people of Taiwan and the Han immigrants from China, they also provide clues to the intricate relationship among the various peoples of Taiwan (Hong, 2002), the transition of rights to land, and the gradual assimilation of the indigenous people (in particular the Pinpu 平埔族群) into the Han society.

Org.	Collection	Number of Land Deeds
NTL	The Archives of the Japanese Taiwan Governor-Generals	15,899
	Published Materials	1,831
	Private Collections	171
NTU	Anli Dashe Archive	2,653
	Published Materials	767
	The Collection of Zheng Family of Xinzhu	383
	Land Deeds from the Department of Anthropology of National Taiwan University	362
	Land Deeds from the Taipei City Archive	153
	Land Deeds of Southern Taiwan	87
TCCC	Published Materials	944
	Total Number of Land Deeds	23,250

Our Collections of Land Deeds of Taiwan

In 2003 and 2004, the Council for Cultural Affairs (CCA) of Taiwan commissioned the National Taichung Library (NTL) and Professor Lee Wen-Liang of NTU to collect and digitize (in full text) the hand-written copies of land deeds from the Archives of the Japanese Taiwan Governor-Generals. In this project, NTL keyed-in the full text of 15,899 land deeds from the Archives of the Japanese Taiwan Governor-Generals. In the meantime, National Taiwan University (NTU) and Taichung County Cultural Center (TCCC) also digitized their own collections of land deeds, most notably the Anli Dashe Archive. Together, NTL, NTU, and TCCC have collected more than 23,000 land deeds in Taiwan, all incorporated into Taiwan History Digital Library (THDL), a full-text digital library of primary historical documents that we built to serve as a research environment for researchers in Taiwanese history and other disciplines. All of the deeds are available in searchable full text, with metadata and, in some cases, images.

The building of content is an on-going effort. We project that the size of our collection will reach 30,000 by the end of the year.

A Research Environment for Land Deeds

We have incorporated the above-mentioned collections of Taiwanese land deeds into Taiwan History Digital Library (THDL) (Chen, Hsiang, Tu, and Wu, 2007), which is built with the goal of providing an electronic research environment for historians. Since our primary goal is to build a digital library to be used by researchers, we spent a great amount of time interacting with historians and built tools that they would find useful in their research. Full-text search is, in our view, the most basic facility. However, what is more important is how to help the user analyze the query results once they are retrieved.

We developed a methodology that treats *query returns as a sub-collection*, instead of as individual (and independent) documents. This seems to reflect better the need of researchers, who usually look for significance emerged from a set of land deeds. Under this philosophy, we have built extensive *post-query classification* facilities, which classify and present the query results according to attributes such as year, type of deeds, origin, etc. We also provide *term frequency analysis* which, using the 50,000 terms (names and locations, mostly extracted automatically) appeared in the collection, analyzes relationships such as geographic locations, co-occurrences, people involved.

The co-occurrence and temporal relationships are further analyzed in the *line chart of temporal distribution* facility provided, which gives a visual representation that makes observation easier. As mentioned before, each land deed features a list of attributes. These attributes can, in principle at least, be extracted from a deed. This work is quite laborious and is still under way. But we have developed an XML format that captures the attributes and, more importantly, makes it easy to build *relation graphs* that show the relationship among land deeds. Our preliminary experiments show that these graphs can play a significant role in the study of land deeds.

In the following, we present the aforementioned features in more detail.

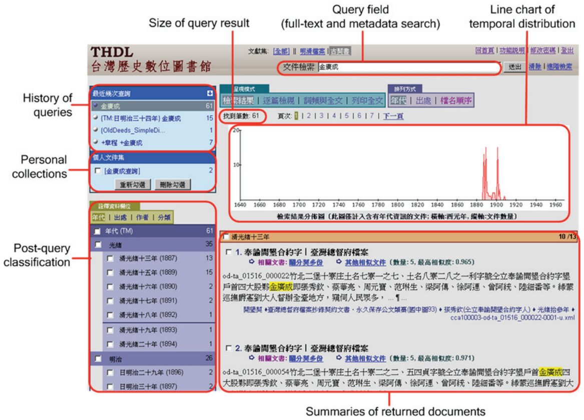
Query Returns as a Sub-Collection

Historians usually do not look at a single document but, rather, a group of documents and try to find significance through their relationship. For example, land deeds from the same region as a whole may reveal the gradual change of land ownership from one ethnic group to another which, obvious, cannot be observed from a single document. For this purpose, we developed a concept that regards the query returns as a *sub-collection* and built tools to help the user find meaning and relationships at the collection level. This is done in THDL mainly through *post-processing* a query's returns, presenting and analyzing them as a whole.

Post-Query Classification

Figure 2 is an example of how query results are presented in THDL. After the query results (the sub-collection) are returned, THDL classifies the resulting land deeds according to three predefined dimensions (year, origin, and type) on the left of the web page, while presenting summaries of the land deeds themselves on the right. Each class is followed by the size of the class (Fig. 3). By representing post-query classification, the historian can observe the *distribution* and behavior of the sub-collection, and see if there is anything that contradicts what the historian predicts. At a first glimpse, the historian can quickly capture the outline of the sub-collection. It's helpful especially when the query results of full-text search is too large to manage. Furthermore, the post-query classification can also be used as a faceted search: simply click on a class will refine the user's query.

Note that the three dimensions are chosen because they are important characteristics of land deeds. A different corpus could define a completely different set of dimensions to reflect the characteristics of the content.



□ 1. 奉諭開墾合約字 | 臺灣總督府檔案
相關文書: 墾分契多份 其他相似文件（數量: 5, 最高相似度: 0.965）

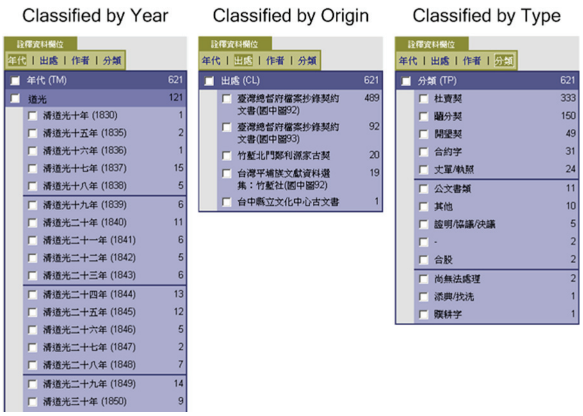
od-ta_01516_000022竹北二堡十寮庄土名七寮一之七、土名八寮二八之一利字號仝立奉諭開墾合約字墾戶首四大股夥金廣成即張秀欽、蔡華亮、周元寶、范琳生、梁阿傳、徐阿連、曾阿統、陸細番等。緣蒙巡撫爵憲劉大人督辦全臺地方，竊伺人民眾多，... ¶...

開墾契 ♦臺灣總督府檔案抄錄契約文書・永久保存公文類纂(國中圖93) ♦ 張秀欽(仝立奉諭開墾合約字人) ♦ 光緒拾參年 ♦
cca100003-od-ta_01516_000022-0001-u.xml

□ 2. 奉諭開墾合約字 | 臺灣總督府檔案
相關文書: 墾分契多份 其他相似文件（數量: 5, 最高相似度: 0.971）

od-ta_01516_000054竹北二堡十寮庄土名十寮二之二、五四貞字號仝立奉諭開墾合約字墾戶首金廣成四大股夥即張秀欽、蔡華亮、周元寶、范琳生、梁阿傳、徐阿連、曾阿統、陸細番等。緣蒙巡撫爵憲劉大

Fig. 2 An outline of THDL right after a query



Classified by Year

詮釋資料欄位
年代 | 出處 | 作者 | 分類

□ 年代 (TM)	621
□ 道光	121
□ 清道光十年 (1830)	1
□ 清道光十五年 (1835)	2
□ 清道光十六年 (1836)	1
□ 清道光十七年 (1837)	15
□ 清道光十八年 (1838)	5
□ 清道光十九年 (1839)	6
□ 清道光二十年 (1840)	11
□ 清道光二十一年 (1841)	6
□ 清道光二十二年 (1842)	5
□ 清道光二十三年 (1843)	6
□ 清道光二十四年 (1844)	13
□ 清道光二十五年 (1845)	12
□ 清道光二十六年 (1846)	5
□ 清道光二十七年 (1847)	2
□ 清道光二十八年 (1848)	7
□ 清道光二十九年 (1849)	14
□ 清道光三十年 (1850)	9

Classified by Origin

詮釋資料欄位
年代 | 出處 | 作者 | 分類

□ 出處 (CL)	621
□ 臺灣總督府檔案抄錄契約文書(國中圖92)	489
□ 臺灣總督府檔案抄錄契約文書(國中圖93)	92
□ 竹塹北門鄭利源家古契	20
□ 台灣平埔族文獻資料選集：竹塹社(國中圖92)	19
□ 台中縣立文化中心古文書	1

Classified by Type

詮釋資料欄位
年代 | 出處 | 作者 | 分類

□ 分類 (TP)	621
□ 杜賣契	333
□ 鬮分契	150
□ 開墾契	49
□ 合約字	31
□ 丈單/執照	24
□ 公文書類	11
□ 其他	10
□ 證明/協議/決議	5
□ -	2
□ 合股	2
□ 尚無法處理	2
□ 添典/找洗	1
□ 贌耕字	1

Fig. 3 Post-query classification according to year, origin, and type of land deeds

Line Chart of Temporal Distribution

The post-query classification on year reviews the temporal distribution of a sub-collection. To better visualize the temporal distribution of a query's returns, we have built a tool to draw a line chart for any given query. It is especially useful when comparing the temporal distributions of two queries at a same time. For example, when a historian suspects that there is dependency between two concepts and wants more analysis, she can simply input each concept as a query, and then get a line chart (Fig. 4). The line chart of Fig 4. suggests that the two concepts are quite correlated.

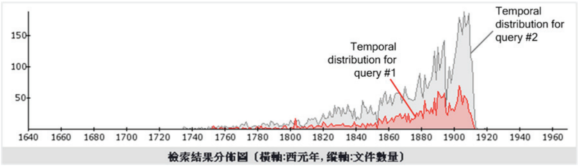
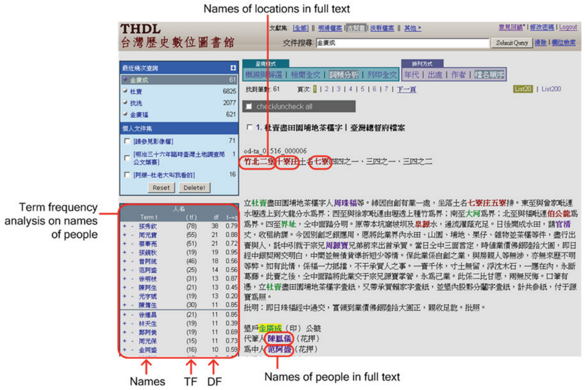


Fig. 4 The temporal distribution of two queries

Term Frequency Analysis

We have developed a term extraction method for extracting noun phrases from old Chinese text (Chang 2006). In the land deed corpus, we have successfully extracted 40,000 names of people and 7,000 names of locations from metadata records and from full text (Chang, 2006). THDL uses the names to provide term frequency analysis by calculating the numbers of times each name appears in the sub-collection and representing the result in tables alongside the full text of resulting land deeds (Fig. 5). The names are listed in descending order according to their document frequency (*DF*, the number of documents in which a name appears). The user can use the tables to observe the relevance among locations and people in the sub-collection. Fig. 5 shows the returns of the query "Jin Guang Cheng" (金廣成), a local reclamation cooperative in the Guanxi (關西) area. At a closer look at the tables of names (Fig. 6) shows that the people on top are indeed the major shareholders of Jin Guang Cheng. Similarly, the locations on top are exactly the locations where Jin Guang Cheng claimed lands back to 1880s. However, the person with the highest DF only appears in 38 documents, while the size of the sub-collection is 61, showing that none of the people appear dominantly in the sub-collection. On the contrary, the locations with the highest DF, "Shiliao Zhuang" (十寮庄) and "Zhubei Er Bao" (竹北二保), appear in 57 and 56 documents accordingly, showing the lands Jin Quang Cheng claimed were mostly around the same area.


Names of locations in full text
Term frequency analysis on names of people
Names
TF
DF
Names of people in full text
THDL
台灣歷史數位圖書館
文獻集: [全部] ‖ 明清檔案 | 古契書 | 淡新檔案 ‖ 其他 ▸
意見回饋 | 修改密碼 | Logout
文件搜尋 金廣成
Submit Query
清除 | 欄位檢索
最近幾次查詢
個人文件集
Reset
Delete!
呈現方式
概觀與篩選 | 檢閱全文 | 詞頻分析 | 列印全文
排列方式
年代 | 出處 | 作者 | 檔名順序
List20 | List200
check/uncheck all

最近幾次查詢	
金廣成	61
杜賣	6825
找洗	2077
金廣福	621

個人文件集	
□ [請參見影像檔]	71
□ [明治三十六年臨時臺灣土地調查局公文類纂]	1
□ [阿繼--杜老大叫我看的]	16

找到筆數: 61 頁次: 1 | 2 | 3 | 4 | 5 | 6 | 7 | 下一頁

□ 1. 杜賣盡田園埔地茶欉字 | 臺灣總督府檔案

od-ta_0_516_000006

竹北二堡十寮庄土名七寮四四之一、三四之一、三四之二

墾戶金廣成（印）公號
代筆人陳凰儀（花押）
爲中人范阿盛（花押）

Fig. 5 An outline of THDL when representing the term frequency analysis facility

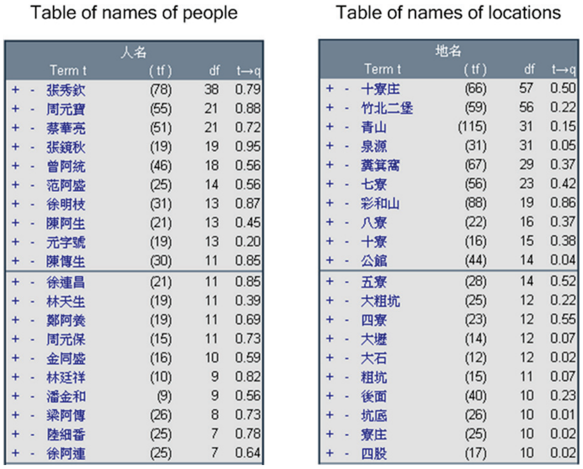


Table of names of people

		人名			
		Term t	(tf)	df	t→q
+	-	張秀欽	(78)	38	0.79
+	-	周元寶	(55)	21	0.88
+	-	蔡華亮	(51)	21	0.72
+	-	張鏡秋	(19)	19	0.95
+	-	曾阿統	(46)	18	0.56
+	-	范阿盛	(25)	14	0.56
+	-	徐明枝	(31)	13	0.87
+	-	陳阿生	(21)	13	0.45
+	-	元字號	(19)	13	0.20
+	-	陳傳生	(30)	11	0.85
+	-	徐連昌	(21)	11	0.85
+	-	林天生	(19)	11	0.39
+	-	鄭阿養	(19)	11	0.69
+	-	周元保	(15)	11	0.73
+	-	金同盛	(16)	10	0.59
+	-	林廷祥	(10)	9	0.82
+	-	潘金和	(9)	9	0.56
+	-	梁阿傳	(26)	8	0.73
+	-	陸細番	(25)	7	0.78
+	-	徐阿連	(25)	7	0.64

Table of names of locations

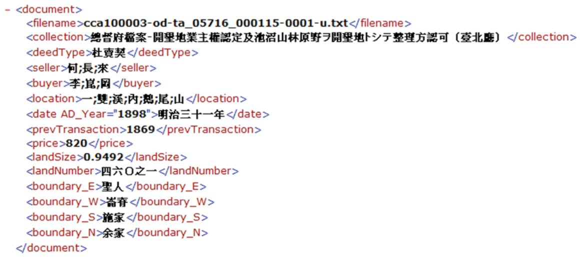
		地名			
		Term t	(tf)	df	t→q
+	-	十寮庄	(66)	57	0.50
+	-	竹北二堡	(59)	56	0.22
+	-	青山	(115)	31	0.15
+	-	泉源	(31)	31	0.05
+	-	糞箕窩	(67)	29	0.37
+	-	七寮	(56)	23	0.42
+	-	彩和山	(88)	19	0.86
+	-	八寮	(22)	16	0.37
+	-	十寮	(16)	15	0.38
+	-	公館	(44)	14	0.04
+	-	五寮	(28)	14	0.52
+	-	大粗坑	(25)	12	0.22
+	-	四寮	(23)	12	0.55
+	-	大壢	(14)	12	0.07
+	-	大石	(12)	12	0.02
+	-	粗坑	(15)	11	0.07
+	-	後面	(40)	10	0.23
+	-	坑底	(26)	10	0.01
+	-	寮庄	(25)	10	0.02
+	-	四股	(17)	10	0.02

Fig. 6 The tables of names

Relation Graphs and Role Analysis

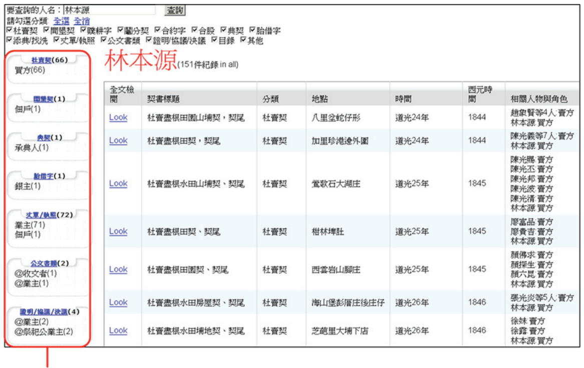
Each land deed should, in principle, include all the attributes we mentioned earlier in the paper. It is thus desirable to extract them so that analysis can be done more easily. We have developed an XML format and have already extracted and analyzed about 13,000 of these deeds (with the attributes of such land deed represented as an XML file). Fig. 7 shows an example of these XML files.

We have also developed a way to show the inter-relationship of the XML files via a notion of relation graphs. These graphs have been used to conduct role analysis, which shows how a specific person, family, or cooperative is involved in the development of a certain region through time. It is done by unfolding the roles they played in land deeds. For example, Fig 8. shows the role analysis of "Lin Benyuan" (林本源), a cooperative that the well-known Lin family of northern Taiwan set up to represent the family in land acquisition. We found that most of the land deeds involving Lin Benyuan in our collection are sales of lands or certificates of lands from the government. Furthermore, Lin Benyuan was the buyer in all the sales and was the landowners in all the certificates. This observation matches the general impression of the Lin family, which has been one of the wealthiest families in Taiwan since late 19th century till now. The timeline of the deeds also shows that they focused on (and systematically) acquiring lands from one geographic location before moving on to the next.



```
- <document>
    <filename>cca100003-od-ta_05716_000115-0001-u.txt</filename>
    <collection>總督府檔案-開墾地業主權認定及池沼山林原野ヲ開墾地トシテ整理方認可（臺北廳）</collection>
    <deedType>杜賣契</deedType>
    <seller>何;長;來</seller>
    <buyer>李;崑;岡</buyer>
    <location>一;雙;溪;內;鵝;尾;山</location>
    <date AD_Year="1898">明治三十一年</date>
    <prevTransaction>1869</prevTransaction>
    <price>820</price>
    <landSize>0.9492</landSize>
    <landNumber>四六〇之一</landNumber>
    <boundary_E>聖人</boundary_E>
    <boundary_W>崙脊</boundary_W>
    <boundary_S>施家</boundary_S>
    <boundary_N>余家</boundary_N>
  </document>
```

Fig. 7 The XML format for representing attributes in land deeds



要查詢的人名：林本源 查詢

請勾選分類 全選 全消

☑杜賣契 ☑間墾契 ☑贌耕字 ☑鬮分契 ☑合約字 ☑合股 ☑典契 ☑胎借字

☑添典/找洗 ☑丈單/執照 ☑公文書類 ☑證明/協議/決議 ☑目錄 ☑其他

杜賣契(66)
買方(66)

間墾契(1)
佃戶(1)

典契(1)
承典人(1)

胎借字(1)
銀主(1)

丈單/執照(72)
業主(71)
佃戶(1)

公文書類(2)
@收文者(1)
@業主(1)

證明/協議/決議(4)
@業主(2)
@祭祀公業主(2)

林本源(151件紀錄 in all)

全文檢閱	契書標題	分類	地點	時間	西元時間	相關人物與角色
Look	杜賣盡根田園山埔契，契尾	杜賣契	八里坌蛇仔形	道光24年	1844	趙衆賢等4人:賣方 林本源:買方
Look	杜賣盡根田契，契尾	杜賣契	加里珍港邊外園	道光24年	1844	陳光義等7人:賣方 林本源:買方
Look	杜賣盡根水田山埔契、契尾	杜賣契	鶯歌石大湖庄	道光25年	1845	陳光賜:賣方 陳光丕:賣方 陳光邦:賣方 陳光波:賣方 陳光清:賣方 林本源:買方
Look	杜賣盡根田契、契尾	杜賣契	柑林埤肚	道光25年	1845	廖富品:賣方 廖貴吉:賣方 林本源:買方
Look	杜賣盡根田園契、契尾	杜賣契	西雲岩山腳庄	道光25年	1845	顏佛求:賣方 顏探生:賣方 顏六昆:賣方 林本源:買方
Look	杜賣盡根水田房屋契、契尾	杜賣契	海山堡彭厝庄後庄仔	道光26年	1846	張光炎等5人:賣方 林本源:買方
Look	杜賣盡根水田埔地契、契尾	杜賣契	芝葩里大埔下店	道光26年	1846	徐妹:賣方 徐露:賣方 林本源:買方

Land deeds are classified into types, with the role analysis of Lin Benyuan

Fig. 8 The role analysis result for querying "Lin Benyuan"

Other Facilities to Assist Research

We have also built other facilities to assist research. THDL allows users to *bookmark* documents and thus create their own sub-collections. Moreover, users can *manipulate sub-collections* by applying three set operations – union, intersection, and complement – on sub-collections and save the result as a sub-collection (Tu, 1998). This facility enables the user to adjust a query's returns by adding or removing designated documents, or documents from another sub-collection. THDL automatically calculates the similarity between documents and reports to user when there are documents similar to the present one. The user can check if it's a duplicate land deed or a copy. For reading assistance, THDL provides the facility highlights query terms and user-designated keywords. This facility can help the user to quickly identify keywords of interests when reading, and thus can help quick evaluation of relevance.

Concluding Remarks

In this paper we described the land deed portion of THDL (Taiwan History Digital Library) that we build with the goal of providing a research environment with primary documents in full-text for research in Taiwanese history. We have developed a concept of regarding query results as a sub-collection, and have built tools that help users observe the relationship and collective meaning of a set of documents. On the aspect of land deeds, we have noticed that most of the existing research in Taiwan on this subject had used no more than a few old deeds (often within a hundred). It would be interesting to see, with over 20,000 land deeds available in searchable full-text and with tools to help discovering and analyzing their relationship, what kind of research issues can emerge and what kind of observations can be made.

Notes

[1] This estimation, however, could be significantly lower than the real number.

Acknowledgements

The authors gratefully acknowledge the National Taichung Library for their effort in producing the full text of many of the land deeds mentioned in this article; the Council of Cultural Affairs of Taiwan for authorizing the inclusion of those deeds into our system; the staff of the Special Collections Division of the National Taiwan University Library in producing the full text of their collection and for further checking the typing correctness of all the deeds; Professors Wu Micha and Lee Wen-Liang of the Department of History of NTU for their continuing support; the members of the Lab of Digital Archives and Automated Reasoning of the Department of Computer Science of NTU for producing many of the tools mentioned here; and the staff of the Research Center for Digital Humanities for their general support.

References

Chang, S.P. (2006). *A Word-Clip Algorithm for Named Entity Recognition: by Example of Historical Documents*. Master Thesis. National Taiwan University.

Chen, S.P., Hsiang, J., Tu, H.C., and Wu, M.C. (2007). On Building a Full-Text Digital Library of Historical Documents. In: Dion Hoe Lian Goh, Tru Hoang Cao, Ingeborg SÃ¸lvberg, Edie Rasmussen (Eds.), *10th International Conference on Asian Digital Libraries (ICADL 2007): Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*. Hanoi, Vietnam, December 10-13, 2007. New York: Springer Berlin, pp. 49-60.

Hong, L.W. (2005). *A Study of Aboriginal Contractual Behavior and the Relationship between Aborigines and Han Immigrants in West-Central Taiwan*. Vol. 1. Taiwan: Taichung County Cultural Center, pp. 5.

Tu, H.C. (1998). *Interactive Web IR: Focalization Model, Effectiveness Measures, and Experiments*. Ph. D. Dissertation. National Taiwan University.

Wu, M.C., Ang K.I., Lee, W.L., and Lin, H.Y. (2004). *A Brief Introduction to the Integrated Collections of Taiwan-related Historical Records*. Taiwan: CCA and Yuan-Liou Publishing, pp. 101.

Wang, S.C. (1977). *Taiwanese Historical Documents in Private Holdings*. Taipei: Huanqiu Publishing.

Wang, S.C. (1993). Introducing Historical Documents of Taiwan: Government Archives, Old Documents, and Genealogies. In: Yan-xian Zhang and Mei-rong Chen eds. *Taiwan History and Historical Documents of Taiwan*. Taipei: Zili Evening News Publishing,

Wang, S.C. (2004). *Papers of Taiwan Historical Documents, Volume 1*. Taipei: Daw Shiang Publishing, pp. 375-376.

Predicting new words from newer words: Lexical borrowings in French

Paula Horwath Chesley
University of Minnesota
ches0045@umn.edu

R. Harald Baayen
University of Alberta
baayen@ualberta.ca

This study models the integration of new lexical borrowings into French, a language in which new lexical borrowings are common. Our goal is to predict whether or not a new lexical borrowing will “survive” the onslaught of time and be integrated into French.

In linguistics, most theories of word formation have been conducted in the generative tradition, such as those taken by Aronoff (1976), Selkirk (1982), Halle & Marantz (1993), and Ussishkin (2005). These approaches work well for new words formed by affixation and address, for example, how to form the neologism *hateable* from hate according to the same rules from which we have *love* --> *loveable*. Yet these theories have not addressed the productivity of borrowings. Although borrowings may have internal morphological structure in the donor language, their adoption in French is not governed by structural rules as studied in theoretical morphology. The goal of the present study is to address the non-structural factors that codetermine whether a borrowing will find its way into the vocabulary of the recipient language.

Although many words from other languages enjoy ephemeral use, the borrowings that become entrenched in the language are a highly constrained subset of the possible borrowings: new words do not occur indiscriminately. Several factors may promote entrenchment in the recipient language’s lexicon.

First, the DONOR LANGUAGE of a borrowing may play a role in lexical integration. For example, borrowings from a prestigious language like English could be more likely to be integrated into the French lexicon than borrowings from a less prestigious language like Polish. Second, a borrowing’s FREQUENCY at a given moment in time could be an influential predictor about the borrowing’s integration into the language at a later point in time. Third, a borrowing’s DISPERSION—the number of different text chunks a word occurs in if a text is divided into several sub-parts—also promises to be a worthwhile predictor. The more writers/speakers use a borrowing, the greater likelihood it has of becoming entrenched in the language community. Fourth, since shorter borrowings require less processing effort, we hypothesize that the LENGTH of the borrowing will be inversely related to the degree of integration of a borrowing. Fifth, the SENSE PATTERN (monosemy or polsemy) of aborrowing in the recipient language may also be at issue. A semantically rich borrowing might have a greater chance of surviving than a semantically and contextually highly restricted, specialized, borrowing. Finally, we consider as well a cultural context factor, whether or not the borrowing refers to a culture that typically corresponds to the language of the borrowing. It is possible that a culturally unrestricted borrowing, for example, a Russian borrowing when describing China, could indicate a greater degree of integration than a restricted cultural context in which a Russian borrowing describes Russia.

	B	S.E.	t value	Pr(>|t|)
(Intercept)	1.496	0.734	2.039	0.042
DISPERSION	2.322	0.123	18.934	<0.001
FREQUENCY	-0.801	1.012	-0.791	0.429
LENGTH	0.599	0.355	-1.688	0.093
CONTEXT (UNRESTRICTED)	0.564	0.833	0.677	0.499
SENSE (POLY)	2.230	0.513	4.347	<0.001
LANGUAGE (ENG)	-0.755	0.530	-1.425	0.155
FREQUENCY* DISPERSION	-3.324	0.692	-4.806	<0.001
FREQUENCY*CONTEXT (UNRESTRICTED)	2.531	0.865	2.927	0.004
LENGTH*CONTEXT (UNRESTRICTED)	-1.721	0.468	-3.676	<0.001
SENSE (POLY)*CONTEXT (UNRESTRICTED)	-2.016	0.744	-2.710	0.007
LANGUAGE (ENG) *CONTEXT (UNRESTRICTED)	1.837	0.586	3.137	0.002

Table1. A multiple regression model for predicting integration of lexical borrowings into French.

This study gathered initial new borrowings from the *Le Monde* corpus (Abeillé *et al*. 2003) from 1989–1992. We alsoqueried the online archives of *Le Figaro* for the borrowings from 1996–2006, taking occurrence in this second corpus as a proxy for integration into the French lexicon. Given the frequency, the dispersion, the length, the donor language of the borrowing, the borrowing's sense pattern and its cultural context in *Le Monde*, we developed a multiple regression model predicting the frequency of occurrence of the borrowing in the later *Le Figaro* corpus. Our model succeeded in explaining a high proportion of the variance in the *Le Figaro* frequencies (R2 =0.673, with minimal overfitting as evidenced by bootstrap validation). Table 1 summarizes this model.

Table 1 shows a highly significant main effect for dispersion. The effect of dispersion is modulated by an interaction with frequency, indicating the role of frequency is restricted to words that have a broad dispersion. Comparing frequency and dispersion, dispersion emerges as the pre-eminent predictor for integration into the lexicon. Length, operationalized in terms of number of syllables, emerged with a negative slope, as expected. The effect of length depended on the cultural context. In culturally unrestricted contexts, longer borrowings are less likely to be integrated into the lexicon. Polysemous borrowings, borrowings with another sense already existing in the language, are also more likely to be integrated into the lexicon than borrowings that do not have another sense. Finally, the cultural context variable turned out to modulate the effect of most other predictors (frequency, length, sense, and donor language).

We have documented a range of factors that codetermine the acceptance of borrowings in a new language. These factors may play a role not only for the entrenchment of borrowings, but also for the entrenchment of regular morphologically complex neologisms, complementing the structure-directed investigations of theoretical morphology. An important direction for future research is to investigate whether, and if so how, the weights of the factors documented here are modulated by the internal structure of complex words (across affixed words, blends, and acronyms).

Unlike other types of word formation, borrowings allow us to gauge the degree of interaction between cultures. The cultural context factor in the present study, for instance, suggests that borrowings can be used to trace how concepts from dominant cultures establish themselves in the language community and spread to those contexts where subordinate cultures are in focus. This information is not only of use to linguists, but also to sociologists and anthropologists.

The methodology outlined in the present study may also be of use for lexicography, as it makes it possible to predict which borrowings (and other neologisms) are in the process of becoming entrenched in the language community, and therefore merit inclusion in dictionaries.

References

Abeillé Clément, & F. Toussenel. 2003. Building a treebank for french. In *Treebanks: Building and Using Parsed Corpora*, 165–188. Kluwer Academic Publishers.

Aronoff, Mark. 1976. Word Formation in Generative Grammar. Cambridge, Mass.: MIT Press.

Halle, M.,& A. Marantz.1993. Distributed morphology and the pieces of inflection. In *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, ed. by K. Hale & S. J. Keyser, volume 24 of Current Studies in Linguistics, 111–176. Cambridge, Mass: MIT Press.

Selkirk, E. 1982. *The Syntax of Words*. Cambridge: The MIT Press.

Ussishkin, A. 2005. A Fixed Prosodic Theory of Nonconcatenative Templatic Morphology. *Natural Language & Linguistic Theory* 23.169–218.

Complementary critical traditions and Elizabeth Cary's *Tragedy of Mariam*

Louisa Connors
University of Newcastle
Louisa.Connors@newcastle.edu.au

It has become a critical commonplace that computational studies have not achieved widespread acceptance in mainstream literary studies (Corns 1991, Potter 1991, Fortier 1991, McGann 2001, Ramsay 2003, Rommel 2004). It is also the case, however, that practitioners have tended to be cautious about the sorts of interpretive claims that can be made on the basis of statistical analysis. Ramsay (2003) locates the source of the caution in a from of "dissonance" in the computing community where empirical studies are seen as providing a kind of safe haven that would be compromised if the researcher were to intervene excessively with either the data or the analysis (168).

The view of computational stylistics as "scientific" and therefore irrelevant to mainstream literary criticism is a pervasive one. Scholars who are sympathetic to the methods involved in attribution and the analysis of style typically construct the problem of humanities computing relationship with the wider scholarly community in terms of a need for computing scholars to rethink their research such that it is framed in terms of "interesting" versus "right" (Ramsay 2003 173) or "interactive tool" versus "quantitative tool" (Sinclair 2003 177) or "disciplined play" versus "unity and coherence" (Rockwell 2003 213). In constructing the debate in these terms there is a risk of amplifying whatever differences there may be between traditional humanities scholars and computing scholars. The result is that the perceived differences become further entrenched. It also sidesteps the question of how and why attribution studies work.

Attitudes to what can be said about a text depend largely on the account of language that underpins an analysis (Connors 2006a; Hoover 2007). Cognitive grammar as developed by Ronald Langacker (1987, 1991) provides a theory that justifies the counting of function words in a computational analysis and also provides an interpretive framework that can explain the use of function words as a rhetorically motivated choice with semantic implications. As such it provides a promising theoretical account for the kinds of computational studies carried out by Burrows (1987a, 1987b, 1992b), Craig (1999a, 2002, 2004), Hoover (2003a, 2003b), Rybicki (2005), and others.

This study explores the use of function words in an analysis of style of 60 tragedies published between 1580 and 1641, with a particular focus on Elizabeth Cary's, *The Tragedy of Mariam*. The set is comprised of the 12 so-called "Sidnean" closet tragedies, as well as 48 tragedies that were written for performance (See Appendix 1). Of these 5 were performed in private theatres and 41 in public theatres. It is not known whether the remaining 2 were performed (Harbage & Schoenbaum, 1964). Some textual preparation was carried out prior to the analysis, but editing and coding was kept to a minimum. Although spelling was regulated homographs were not tagged. Contracted forms throughout the texts were expanded so that their constituents appeared as separate words.

Tagged texts were then run through a frequency count using a program developed by the Centre for Literary and Linguistic Computing (CLLC) at the University of Newcastle called *Intelligent Archive* (IA). A list of 241 function words was used (see Appendix 2). This list was developed on the basis of work by the CLLC. Using IA a frequency test was run to establish which function words were most commonly found in each of the plays. The frequency count for each word in the function word list was expressed in the IA output as a percentage of the total dialogue in the relevant play. The data produced by IA were then transferred to *Statistical Package for the Social Sciences (SPSS),* and Excel, for further analysis.

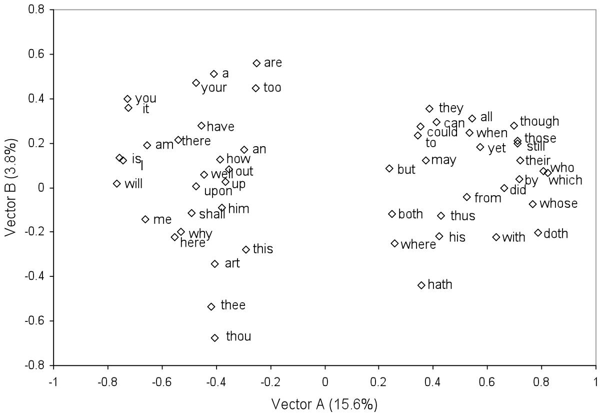


Figure 1. Principal component analysis for 60 tragedies (1580-1641) in 4000 word segments for 53 most discriminating function words – word plot for first two eigenvectors

On the basis of strong discriminate analysis results an independent samples *t* test was carried out to identify a set of "closet" and "stage" markers (see appendix 3). The markers were only selected if they also satisfied both the *t* test for equality of means and *Mann Whitney* test. A Principal Component Analysis (PCA) was then carried out to identify which variables were responsible for most of the difference between the sets of texts. For the PCA, the texts were broken into 4000 word segments to ensure that results could be easily presented. Segments of less than 4000 words were incorporated into the preceding segment. This produced 268 segments in total (50 closet segments and 218 stage play segments). The results of the PCA are set out in Table 1 and Table 2.

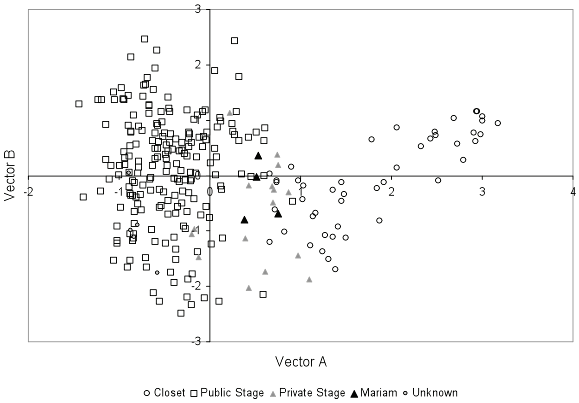


Figure 2. Principal component analysis for 60 tragedies (1580-1641) in 4000 word segments for 53 most discriminating function words

One of the most interesting features of Figure 2 is the way the graph identifies three loosely clustered generic groupings. These are closet plays to the east, public stage plays to the west, and private stage plays along with *Mariam* in the centre. Marta Straznicky (2004) analyses *Mariam* in the context of drama associated with the private theatres. She notes that writers for the stage actively incorporated elements of the neo-Senecan tradition associated with closet tragedy into their plays. While some writers for the stage adopt various features commonly found in closet tragedy to a greater extent than their colleagues, Elizabeth Cary appears to draw on a more theatrical tradition than her closet contemporaries.

This is an interesting result for Elizabeth Cary's play because it supports the insights of more traditional humanities scholars who have described *Mariam* as the most "theatrical" of the Sidnean closet tragedies (Barish 1993; Findlay, Williams & Hodgson-Wright 1999; Straznicky 2004). The play features sub-plots, stage directions and action that is played out rather than dealt with through narrative. Function word analysis supports the view of *Mariam* as one of the more theatrical of the closet texts, and provides additional information about how *Mariam* relates to contemporary texts in terms of the selected variables.

One of the strongest differences between the plays written for the stage, and closet plays, is the use of pronouns. There is evidence of frequent usage of the first and second person singular pronouns in the stage plays and little evidence of their use in closet plays. Van Hoek (2007) has analysed pronouns from the perspective of Langacker's cognitive grammar, particularly in relation to the way they influence the accessibility of concepts and the viewpoint from which constructions are understood. Van Hoek's analysis can usefully explain aspects of closet tragedy and *The Tragedy of Mariam* which have caused readers over time to find the text(s) less engaging than plays written for the public stage.

In a cognitive analysis, language that features frequent use of pronouns, particularly anaphoric pronouns, (such as *him* in "I'll go and get him") assumes a certain common ground between the speaker and the addressee; they both understand who or what is being referred to without "him" being referred to by name. When a full-noun is used, like a name, there is an implied conceptual distance. The notion of "conceptual distance" also relates to how easily accessible a concept is for the discourse participants.

In her first soliloquy, a speech of seventy-five lines, Mariam refers to herself by name four times. She asks herself "then why grieves Mariam Herod's death to hear?" (I.i.38) and calls herself "heard-hearted Mariam" (I.i.62). In most situations it is unusual for a speaker to refer by name to himself or herself. When Mariam names herself, she places the conception of "Mariam" on the metaphorical stage shared between speaker and addressee. In cognitive linguistic terms, it is as though Mariam is observing herself from an outside position. It's a device that breaks up the normal relationship between speaker and addressee and has a disorienting effect.

In his analysis of the Senecan tradition in England, Braden asks "what makes it appropriate for a character to talk this way?" (1985 66). There could be a number of reasons for Cary's characters to engage in this type of self-referential linguistic behaviour, however it is likely that the style of speech is rhetorically motivated. Although the language of the play is somewhat alienating, it reflects serious philosophical, cultural and political ambitions. *Mariam* is explicitly concerned with the individual's right to resist tyranny. The form this resistance takes reflects the influence of neo-Stoic discourses and Seneca as literary and philosophical model.

For Braden the characteristic feature of Senecan drama is a kind of "cosmic self-dramatization" (178) that is represented in the characters by a quest for "a radical, unpredicated independence" (67). The speech of Cary's characters can be linked to this very neo-Senecan idea of selfhood and the neo-Stoic ideals of control of the self, especially in the face of unchecked abuse of power. Mariam's conquest of self can be traced through the way she overcomes emotional confusion to become "free, lofty, fearless and steadfast" (Seneca qtd Straznicky 1994 115). Evidence of this struggle in *Mariam* can be seen at multiple levels, including the linguistically at the level of function words.

References

Barish, Jonas. "Language for the Study: Language for the Stage." The Elizabethan theatre XII. Eds. A. L. Magnusson and C. E. McGee. Toronto: P.D. Meany, 1993. 19-43.

Braden, Gordon. Renaissance Tragedy and the Senecan Tradition: Anger's Privilege. New Haven: Yale University Press, 1985.

Burrows, J.F. Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. New York: Oxford University Press, 1987a.

---. "Computers and the Study of Literature." Computers and Written Texts. Ed. C.S. Butler. Oxford: Blackwell, 1992b.

---. "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style." Literary and Linguistic Computing 2 (1987b): 61-70.

Cary, Elizabeth (Lady Falkland). The Tragedy of Mariam the Faire Queene of Jewry. London: Richard Hawkins, 1613.

Connors, Louisa. "Linking Cognitive Linguistics and Computational Stylistics." ALLC Digital Humanities. Sorbonne Paris IV: Centre Culture Anglophones et Technologies de l'Information (CATI), 2006a. 46-47.

Corns, Thomas N. "Computers in the Humanities: Methods and Applications in the Study of English Literature." Literary and Linguistic Computing 6.127-30 (1991).

Craig, D.H. "Common-Word Frequencies, Shakespeare's Style, and the *Elegy* by W.S." Early Modern Literary Studies 8.1 (2002): 1-42.

Craig, Hugh. "Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You

Learned Anything About Them.” Literary and Linguistic Computing 14.1 (1999a): 103-12.

---. “Stylistic Analysis and Authorship Studies.” A Companion to Digital Humanities. Eds. Susan Schreibman, Ray Siemens and John Unsworth: Blackwell, 2004. 273-88.

Findlay, Alison, Gweno Williams, and Stephanie Hodgson-Wright. “”The Play Is Ready to Be Acted”: Women and Dramatic Production, 1570-1670.” Women’s Writing 6.1 (1999): 129-48.

Fortier, P.A. “Theory, Methods and Applications: Some Examples in French Literature.” Literary and Linguistic Computing 6.192-6 (1991).

Harbage, Alfred, and S. Schoenbaum, eds. Annals of English Drama 975-1700: An Analytical Record of All Plays, Extant or Lost, Chronologically Arranged. 1940. Second ed. Philadelphia: University of Philadelphia Press, 1964.

Hoover, D.L. “Another Perspective on Vocabulary Richness.” Computers and the Humanities 37 (2003b): 151-78.

---. “The End of the Irrelevant Text.” Digital Humanities Quarterly 1.2 (2007).

Hoover, David L. “Multivariate Analysis and the Study of Style Variation.” Literary and Linguistic Computing 18.4 (2003a): 341-60.

Langacker, Ronald W. Foundations of Cognitive Grammar: Descriptive Application. Vol. II. Stanford, California: Stanford University Press, 1991.

---. Foundations of Cognitive Grammar: Theoretical Prerequisites. Vol. I. Stanford, California: Stanford University Press, 1987.

McGann, Jerome. Radiant Textuality: Literature after the World Wide Web. New York: Palgrave, 2001.

Potter, R.G. “Statistical Analysis of Literature: A Retrospective.” Computers and the Humanities 25.401-29 (1991).

Ramsay, Stephen. “Toward an Algorithmic Criticism.”

Author	Title	Year first presented	Copytext	Date of publication of earliest text	Date of publication of copytext	Classification	Auspcies	Status
Note: Authorship, date, date of publication of earliest text, classification and auspices are as in Harbage (1964) *Annals of English Drama, second edition*								
Kyd, Thomas	*The Spanish Tragedy*	1587	STC15086	1592	1592	Tragedy	Strange's	Public
Marlowe, Christopher	*The Jew of Malta*	1589	STC17412	1633	1633	Tragedy	Strange's	Public
Sidney(Herbert), Mary	*Antonius*	1590	STC11623	1592	1595	Tragedy	Closet	Closet
Anonymous	*Arden of Faversham*	1591	STC733	1592	1592	Realistic Tragedy	Unknown	Unknown
Wilmot, Robert, et al.	*Tancred and Gismund*	1591*	STC25764	1591	1591	Senecan Tragedy	Inner Temple	Private
Marlowe & Rowley	*Doctor Faustus*	1592	STC17432	1604	1616	Tragedy	Admiral's (by 1594)	Public
Daniel, Samuel	*Cleopatra*	1593	STC6254	1594	1594	Tragedy	Closet	Closet
Kyd, Thomas	*Cornelia*	1594	STC11622	1594	1594	Tragedy	Closet	Closet
Shakespeare, William	*Titus Andronicus*	1594	STC22328	1594	1594	Tragedy	Pembroke's/Sussex's	Public
Shakespeare, William	*Romeo and Juliet*	1595	STC22323	1597	1597	Tragedy	Chamberlain's	Public
Greville, Fulke	*Mustapha*	1596	STC12362	1609	1609	Tragedy	Closet	Closet
Brandon, Samuel	*The Virtuous Octavia*	1598	Malone Soc. Reprint	1598	1598	Tragicomedy	Closet	Closet
Shakespeare, William	*Julius Caesar*	1599	STC22273	1623	1623	Tragedy	Chamberlain's	Public
Marston, John	*Antonio's Revenge*	1600	STC17474	1602	1602	Tragedy	Paul's	Private
Greville, Fulke	*Alaham*	1600		1633	1633	Tragedy	Closet	Closet
Shakespeare, William	*Hamlet*	1601	STC22276	1603	1604	Tragedy	Chamberlain's	Public
Shakespeare, William	*Troilus and Cressida*	1602	STC22331	1609	1609	Tragedy	Chamberlain's	Public
Heywood, Thomas	*A Woman Killed with Kindness*	1603	STC13371	1607	1607	Tragedy	Worcester's	Public
Jonson, Ben	*Sejanus His Fall*	1603	STC14782	1605	1605	Tragedy	King's	Public
Alexander, William	*Darius*	1603		1603	1637	Tragedy	Closet	Closet
Cary, Elizabeth	*Mariam, The Fair Queen of Jewry*	1604	STC4613	1613	1613	Tragedy	Closet	Closet
Shakespeare, William	*Othello*	1604	STC22305	1622	1622	Tragedy	King's	Public
Alexander, William	*Croesus*	1604		1604	1637	Tragedy	Closet	Closet
Daniel, Samuel	*Philotas*	1604		1605	1605	Tragedy	Queen's Revels	Closet
Marston, John	*The Wonder of Women*	1605	STC17488	1606	1606	Tragedy	Queen's Revels	Private
Shakespeare, William	*King Lear*	1605	STC22292	1608	1608	Tragedy	King's	Public
Anonymous (Tourneur? Middleton?)	*The Revenger's Tragedy*	1606	STC24150	1607 or 08	1608	Tragedy	King's	Public
Anonymous (Shakespeare; Wilkins?)	*A Yorkshire Tragedy*	1606	STC22340	1608	1608	Tragedy	King's	Public
Shakespeare, William	*Macbeth*	1606	STC22273	1623	1623	Tragedy	King's	Public

Appendix 1. List of plays used in this study

Literary and Linguistic Computing 18.2 (2003): 167-74.

Rockwell, Geoffrey. “What Is Text Analysis, Really?” Literary and Linguistic Computing 18.2 (2003): 209-19.

Rommel, Thomas. “Literary Studies.” A Companion to Digital Humanities. Eds. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell, 2004.

Rybicki, Jan. “Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz’s Trilogy.” Literary and Linguistic Computing 26.1 (2005): 91-103.

Sinclair, Stefan. “Computer-Assisted Reading: Reconceiving Text Analysis.” Literary and Linguistic Computing 18.2 (2003): 175-84.

Straznicky, Marta. Privacy, Play Reading, and Women’s Closet Drama, 1550-1700. New York: Cambridge University Press, 2004.

---. “Profane Stoical Paradoxes: *The Tragedy of Mariam* and Sidney Closet Drama.” English Literary Renaissance 24 (1994): 104-34.

van Hoek, Karen. “Pronominal Anaphora.” Oxford Handbook of Cognitive Linguistics. Eds. Dick Geeraerts and Hubert Cuyckens. Oxford: Oxford University Press, 2007. 890-915.

a
about
above
across
after
again
against
all
almost
along
also
although
always
am
amid
amidst
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anywhere
are
around
art
as
at
be
because
been
before
behind
being
below
beneath
beside
besides
between
beyond
both
but
by
can
cannot
canst
could
couldst
despite
did
do
does
doing
done
dost
doth
down
during
each
either
enough
even
ever
every
everybody
everyone
everything
everywhere
few
first
for
from
had
has
hast
hath
have
having
he
hence
her
here
hers
herself
him
himself
his
how
however
I
if
in
inside
instead
into
is
it
its
itself
just
least
less
like
little
many
may
mayst
me
might
mightst
mine
more
most
much
must
my
myself
neither
never
nevertheless
no
nobody
none
noone
nor
not
nothing
nowhere
of
off
often
on
one
only
onto
opposite
or
other
ought
oughtst
our
ours
ourselves
out
outside
over
own
past
perhaps
provided
quite
rather
round
same
several
shall
shalt
she
should
shouldst
since
so
some
somebody
somehow
someone
something
somewhere
soon
still
such
than
that
the
thee
their
theirs
them
themselves
then
there
thereupon
these
they
thine
this
those
thou
though
through
throughout
thus
thy
thyself
till
too
to
towards
under
underneath
unless
until
unto
up
upon
us
very
was
we
well
were
what
whatever
when
whenever
where
wherever
which
while
who
whom
whose
why
will
wilst
wilt
within
without
with
would
wouldst
yet
you
your
yours
yourself
yourselves

Appendix 2. List of function words.

Shakespeare, William	*Antony and Cleopatra*	1607	STC22273	1623	1623	Tragedy	King's	Public
Shakespeare, William	*Timon of Athens*	1607	STC22273	1623	1623	Tragedy	Unacted?	Unknown
Alexander, William	*Julius Caesar (WA)*	1607		1607	1637	Tragedy	Closet	Closet
Alexander, William	*The Alexandrean Tragedy*	1607		1607	1637	Tragedy	Closet	Closet
Shakespeare, William	*Coriolanus*	1608	STC22273	1623	1623	Tragedy	King's	Public
Tourneur, Cyril	*The Atheist's Tragedy*	1609	STC24146	1611 or 12	1611	Tragedy	King's	Public
Beaumont and Fletcher	*The Maid's Tragedy*	1610	STC1676	1619	1619	Tragedy	King's	Public
Chapman, George	*The Revenge of Bussy d'Ambois*	1610	STC4989	1613	1613	Tragedy	Queen's Revels	Private
Anonymous (Middleton)	*The Second Maiden's Tragedy*	1611	Malone Soc. Reprint	MS	1611	Tragedy	King's	Public
Jonson, Ben	*Catiline his Conspiracy*	1611	STC14759	1611	1611	Tragedy	King's	Public
Webster, John	*The White Devil*	1612	STC25178	1612	1612	Tragedy	Queen Anne's	Public
Fletcher, John	*Valentinian*	1614	Wing B1581	1615	1647	Tragedy	King's	Public
Webster, John	*The Duchess of Malfi*	1614	STC25176	1623	1623	Tragedy	King's	Public
Goffe, Thomas	*The Courageous Turk*	1618	STC11977	1632	1632	Tragedy	Christ Church, Oxford	Private
Middleton, Thomas	*Hengist, King of Kent*	1618	ed. R. C. Bald	MS	1618**	Tragedy	Unknown (King's 1541)	Public
Fletcher, John	*The Bloody Brother*	1619	STC11064	1639	1639	Tragedy	King's (?)	Public
Rowley, William	*All's Lost by Lust*	1619	STC21425	1633	1633	Tragedy	Prince's (Lady Elizabeth's)	Public
Fletcher and Massinger	*The Double Marriage*	1620	Wing B1581	1647	1647	Tragedy	King's	Public
Middleton, Thomas	*Women Beware Women*	1621	Wing M1989	1657	1657	Tragedy	King's (?)	Public
Markham and Sampson	*Herod and Antipater*	1622	STC17401	1622	1622	Tragedy	Red Bull Company (Revels)	Public
Middleton and Rowley	*The Changeling*	1622	Wing M1980	1653	1653	Tragedy	Lady Elizabeth's	Public
Massinger, Philip	*The Unnatural Combat*	1626	STC17643	1639	1639	Tragedy	King's	Public
Massinger, Philip	*The Roman Actor*	1626	STC17642	1629	1629	Tragedy	King's	Public
Ford, John	*The Broken Heart*	1629	STC11156	1633	1633	Tragedy	King's	Public
Shirley, James	*The Traitor*	1631	STC22458	1635	1635	Tragedy	Queen Henrietta's	Public
Shirley, James	*Love's Cruelty*	1631	STC22449	1640	1640	Tragedy	Queen Henrietta's	Public
Ford, John	*Love's Sacrifice*	1632	STC11164	1633	1633	Tragedy	Queen Henrietta's	Public
Ford, John	*Tis Pity She's a Whore*	1632	STC11165	1633	1633	Tragedy	Queen Henrietta's	Public
Suckling, John	*Aglaura*	1637	STC23420	1638	1638	Tragedy	King's	Public
Davenant, William	*The Unfortunate Lovers*	1638	Wing D348	1643	1643	Tragedy	King's	Public
Shirley, James	*The Cardinal*	1641	Wing S3461	1653	1652	Tragedy	King's	Public

* Harbage notes that *Gismond of Salerne* by Wilmot.R; Stafford; Hatton; Noel; Al., G. was performed in 1566-1568 and revised by Wilmot in 1591 as *Tancred and Gismond*

** Edited from the manuscript in the Folger Shakespeare Library and published 1938

Appendix 1 Continued. List of plays used in this study.

Rank	Orig. order	Word-variable	t	df	Mean Difference	Std. Error Difference	Significance *t* test	Significance Mann
1	16	you	-17.78	142	-1.17	0.07	0	0
2	24	it	-18.38	122	-0.86	0.05	0	0
3	38	will	-16.42	106	-0.52	0.03	0	0
4	126	am	-13.38	162	-0.18	0.01	0	0
5	8	I	-12.3	101	-1.15	0.09	0	0
6	32	me	-11.43	114	-0.45	0.04	0	0
7	176	**doth**	14.64	54	0.34	0.02	0	0
8	114	there	-11.07	124	-0.14	0.01	0	0
9	42	your	-10.93	108	-0.46	0.04	0	0
10	123	here	-9.22	266	-0.18	0.02	0	0
11	18	is	-10.91	64	-0.81	0.07	0	0.0001
12	29	**with**	8.5	266	0.29	0.03	0	0
13	156	upon	-9.23	102	-0.11	0.01	0	0
14	86	**which**	10.58	56	0.37	0.04	0	0
15	108	**yet**	9.84	62	0.20	0.02	0	0
16	89	**from**	8	266	0.15	0.02	0	0
17	162	why	-8.47	99	-0.10	0.01	0	0
18	170	up	-7.9	99	-0.08	0.01	0	0
19	192	**whose**	8.44	55	0.15	0.02	0	0
20	75	shall	-6.98	266	-0.18	0.03	0	0
21	138	**who**	8.44	50	0.39	0.05	0	0
22	58	him	-7.36	99	-0.23	0.03	0	0
23	131	**can**	6.83	266	0.11	0.02	0	0
24	180	**still**	7.97	54	0.16	0.02	0	0
25	150	well	-7.12	102	-0.09	0.01	0	0
26	194	art	-6.94	109	-0.06	0.01	0	0
27	11	a	-6.51	266	-0.46	0.07	0	0
28	130	**did**	7.7	51	0.37	0.05	0	0
29	84	**their**	7.27	60	0.30	0.04	0	0
30	182	**those**	6.95	54	0.15	0.02	0	0
31	92	thee	-6.16	96	-0.17	0.03	0	0
32	186	**though**	6.57	52	0.18	0.03	0	0
33	6	**to**	6	105	0.30	0.05	0	0
34	72	**by**	6.46	52	0.32	0.05	0	0.0001
35	151	**hath**	5.45	266	0.09	0.02	0	0
36	60	thou	-5.6	94	-0.24	0.04	0	0
37	163	an	-5.19	266	-0.06	0.01	0	0
38	35	**but**	5.18	266	0.15	0.03	0	0

Appendix 3. Highly significant differentiations (closet markers in bold) among the 100 most common function words in 60 tragedies (1580-1641) in 4000 word segments: variables that satisfy both t test for equality of means and Mann Whitney test.

The 385+ Million Word *Corpus of Contemporary American English* (*1990-present*): A new tool for examining language variation and change

Mark Davies
Brigham Young University
mark_davies@byu.edu

The last 15-20 years have seen the introduction of "mega-corpora" such as the Bank of English and the British National Corpus, which contain anywhere from 100-500 million words. Until recently, however, there have been no large, balanced corpora of American English. Only a small portion of the Bank of English, for example, is from the US. The well-known *American National Corpus* has not been updated in several years, it has only 22 million words of text, and it is quite unbalanced in terms of genre representation (essentially no fiction, no popular magazines, etc). On the other hand, there are other large "corpora" of American English (such as the GigaWord collection of newspaper articles), but these represent just one genre.

The situation has changed recently, with the recent introduction of the *Corpus of Contemporary American English (COCA)* (www.americancorpus.org), which we released in Spring 2008. This is the first large, balanced corpus of American English, and it will permit researchers to address many questions related to language change and linguistics variation, which could not have been answered until this time. The corpus is composed of more than 385 million words of text in more than 150,000 articles and books, with at least 20 million words each year from 1990 to 2008 (and it will be updated from this point on as well). In each year, the corpus is divided into five equally-sized genres: spoken (transcripts of unscripted conversation on 100+ TV and radio programs each year), fiction (novels, short stories, and movies scripts), popular magazines (100+ magazines each year), newspapers (ten newspapers from across the US), and academic journals (100+ journals each year). The wide range of genres means that researchers can study in detail variation between these genres, and the consistency in genres across time means that researchers can accurately study linguistic changes. In addition, the corpus is tagged and lemmatized (using CLAWS, the same tagger that was used for the British National Corpus), which greatly facilitates syntax-oriented queries.

The entire corpus architecture and interface are designed to facilitate research into language variation and change. Users can quickly and easily find the frequency of any word, phrase, substring (e.g. suffixes), or syntactic construction in each year since 1990, and in each of the five major registers. Example might be words such as *carbon-neutral* or the quotative *like*, phrases like *perfect storm* or *tipping point*, suffixes like *–gate* (*Iraqgate* or *zippergate*), or grammatical constructions like preposition stranding, zero relative clauses, or the 'get passive'. They can also see detailed information on frequency and distribution of words and constructions in micro-genres, such as the rise of *bling* in African-American and entertainment-related popular magazines.

The corpus also allows users to find the collocates in different genres and groups of years since 1990, which can provide valuable insight into semantic change and variation. For example, they can compare the collocates of *woman* or of *peace* in spoken, fiction, and newspapers to see how these concepts are viewed and discussed differently in the two genres. They can also compare collocates over time, such as the increasingly positive collocates with geek since the early 1990s, or the increasing environmental emphasis over time, evidenced by new collocates with green. Finally, they can easily compare the collocates of two words to see contrasts in the meaning or usage of the two terms. Examples might be adjectives with Democrats vs. Republicans (*electable*, *fun*, *open-minded* vs. *extremist*, *mean-spirited*, and *greedy*), or bias in the collocates with *women* and *men* (*glamorous*, *real-life*, *disadvantaged* vs. *honorable*, *self-made*, and *wise*).

Other features allow for fairly complex semantically-oriented searches. Due to the relational database architecture, we have been able to integrate a thesaurus with entries for 60,000+ headwords, as well as WordNet, and users can also create "customized lists" on the fly. These allow for rather powerful queries, such as "any form of any synonym related to the verb *clean* within five words of any word in the 'houseItems' list created by Jones" (*clean the pots, washing some windows, the floor he mopped*) or "any hyponym of emotions within five words of a word in the 'familyTerms' list created by Smith (*Grandpa seemed to be pretty happy, the excited children, the moms that are most worried*). As can be seen, this goes far beyond most other corpus architectures, which are often limited to just word, phrase, lemmas, or parts of speech.

This example of semantically-oriented searches leads us finally to a brief discussion of the overall challenge of designing an architecture that achieves the three competing goals of 1) size 2) speed, and 3) annotation. Achieving two of three goals is relatively simple, but achieving all three simultaneously – in the real world – is much more difficult. For example, there are many search engines that allow fast retrieval from very large "corpora" or text archives (e.g. Google or Lucene), but which allow for little if any annotation (e.g. even basic part of speech tagging or lemmatization, much less integration with thesauruses or user-defined lists). Other approaches provide speed and annotation, but are completely inadequate in terms of scalability – either in terms of size and/or speed. There is no limit to the number of proprietary architectures that have been designed over the past decade or two, and which might work very well for a small one million word corpus, but which are utterly unscalable. A query might take just two or three seconds for a well-annotated 10 million word corpus, but (assuming linearity), that same query then takes two minutes or more for a 350-400 million word corpus.

Our approach – which is based on a (still) proprietary architecture involving relational databases and a massively-redundant n-grams schema – is one of just a handful that adequately allows for size, speed, and annotation. Even a complicated query– involving part of speech, lemma, synonyms, customized word lists, and limited by sub-genres – typically takes just 2-3 seconds to generate results from the entire 385+ million word corpus. In addition, ours is the only architecture (as far as we are aware) that allows for such a wide range of comparisons – e.g. across sub-corpora, or the collocates of different words. For example, SketchEngine allows comparisons between different words, but not by sub-corpus. The IMS Corpus Workbench (CWB) allows comparisons between sub-corpora, but not between different words. Ours offers both of these, full integration with thesauruses and lexical resources like WordNet, as well as much more.

In summary, the Corpus of Contemporary American English (COCA) is based on an architecture and interface that allows for a wide range of queries, and which does so quickly and easily. In terms of the textual database, it is both large (385+ million words, and growing) and well-balanced (in terms of genres and sources). All of these features serve to create a unique resource that allows researchers to look at a wide range of questions dealing with recent changes and current variation in American English, which would have been difficult or impossible to investigate before this time.

The « Bibliothèques Virtuelles Humanistes » (Virtual Humanistic Libraries in Tours): a Collection, or a Corpus?

Marie Luce Demonet
Institut Universitaire de FranceCentre d'Etudes Supérieures de la Renaissance (CESR)
marie-luce.demonet@univ-tours.fr

1. The goal of the Bibliothèques Virtuelles Humanistes (BVH, or Virtual Humanist Libraries, a digitization project begun in Tours in 2003, http://www.bvh.univ-tours.fr) is to offer two types of digital representations of a selection of books printed during the Renaissance or of manuscripts: the image of a copy (its "facsimile") on the one hand, and its transcription on the other hand, without additions besides corrections or variations that are essential for understanding the text, and TEI encoding. These two goals necessitate the combined efforts of two communities whose objectives, methods and formulations are very different. The difference is further complicated as the elaboration of a corpus for this period presents additional difficulties: librarians and book historians work with "image processing" computer programs, whereas literature and language specialists use linguistic systems (Hyperbase, Weblex, Philologic...). Currently, technical progress has allowed these different approaches to come together, although they have not yet been combined: libraries prefer to offer text-only versions of documents, obtained via OCR or by manual transcription encoded in order to show a readable text alone or along with its facsimile. Linguistic corpora and text databases for works before 1800 are often constituted of modern editions, which are under copyright and impossible to show next to "their" facsimiles—which often do not exist, as they were established from several different reference editions and do not respect the physical presentation. These editions have the obvious advantage of easily lending themselves to searches for data and to detailed encoding. The goal of libraries of text images is entirely different, as research is generally done only on metadata and, at best, on the table of contents which constitutes a minimal indexation, and sometimes is done through a quick round of OCR. In light of these two traditions, and taking into account the reading habits and requests that recent navigational tools have encouraged, the BVH are devoted to conducting research on the two-pronged front of indexing text in image mode, extracting images from images of the pages, classifying and indexing them, and acquiring a significant corpus of transcribed texts. But does the collection of texts constitute a corpus?

2. It is revealing that this proposition of intervention is on the fence between the two initial themes of the congress and we defend the name of "corpus" for the image-and-text combination, although it was imposed upon us, so as not to frighten financiers and the general public, to call it a "library". We will discuss the difficulties that this grouping engendered, difficulties both theoretical and technical. The digital libraries as they are comprised at the CESR would not be a "corpus", but rather a "collection" as their only commonality is their period of publication, from 1470-1650, identified as the "Renaissance" in the largest sense, including Antique or Medieval texts edited during that period. Even if the collection includes broad categories (classics of the Renaissance—sources of science—legal and political history—philosophy and theology), *a priori* each book on any shelf could fall under one of these categories, all the more as a fifth category, "particular projects,“ allows additions to one or several subcategories (like the "Rabelais" database, the dictionaries, etc.), the only ones in fact that merit the name of "corpus", the rest being composed of standalone books or curiosities. This is how the members of the jury of the National Research Agency (ANR) understood the word "corpus" in France, as it is obvious from the projects chosen that the notion was relatively vague, even as they excluded a collection of digitized texts in image mode alone. In order for there to be a corpus, there must be text.

3. Going against this analysis, we would defend:

3. 1. The idea that each work chosen corresponds to an analysis of its form and its content. The researchers who are charged with this (under the supervision of Toshinori Uetani, CESR, with the collaboration of Marie-Elisabeth Boutroue, IRHT) examine its interest from the point of view of the history of the book and of the directions of research that obviously reflect the options of researchers at the Center or their colleagues. They hope to render the object upon which they are working accessible to the scientific community, in order to share knowledge, going against the traditional editorial process that consists of offering a "definitive" paper edition once the establishment has finished it. Selection is therefore a tool of anonymous collaboration and results from a step that is one of a researcher, using available funds (as it happens, those of the Région Centre and its partner establishments) that they enrich in so doing: the library of the Museum of Sologne, in Romorantin, possesses a copy of the 1580 edition of Montaigne's *Essais*: although it is not extraordinarily rare, this state of the text merited being offered to the public in order to compare it with that of the BnF's

copy (Gallica).

3. 2. The exploitation of image mode, in particular illustrated elements, is specific to the digitized collection: the AGORA program, developed by the computer sciences laboratory of the University of Tours, allows semi-automatic extraction of illustrations, graphs, portraits, typographic material and initials, providing particular indexed and searchable databases(notably thanks to Thesaurus Iconclass). These "processed" works assuredly constitute a corpus which can be subject to precise searches.

4. As for the texts, the standard is to accept the name of "corpus" for a database that gathers together transcriptions of Renaissance texts in French.

4. 1. But what about texts in Latin, Italian, English, Spanish, etc., that are not excluded *a priori*? A multilingual corpus is all the more conceivable as many Renaissance texts contain large fragments of text in Latin (the *Essais*, for example). Even if we initially preferred texts in French, originals or translations, leaving to researchers from other linguistic regions the care of developing their own corpora, the fluidity of TEI encoding allows the possibility of making a "Renaissance" corpus that would not be limited to French and would contain even bilingual or multilingual "alined" corpora. Thanks to the "TEI Renaissance and modern times" application (Tours, July 2008), appropriate encoding of different versions of a text makes a model available that renders the physical description of a text compatible with its logical description (Nicole Dufournaud, CESR, Jean-Daniel Fekete, INRIA). These procedures are taught in a specialized Master's program and during workshops open to students, researchers and librarians.

4. 2. The progress of "Renaissance" OCR (RETRO) also developed in Tours by Jean-Yves Ramel allows the acquisition of text from difficult-to-read printed matter, and it allows correction thanks to form dictionaries compiled from transcriptions and in different languages. Even if acquisition in text mode with an accuracy rate of over 97% still represents a considerable cost for these early printed books (post-correction is always necessary), it allows incrementation of a collection of texts with highly varying written forms, processed by an annotation tool (Analog, developed in Poitiers and at the ENS-LSH by Marie-Hélène Lay) or a "dissimilation" tool (Thierry Vincent, Poitiers); these collections in turn facilitate the acquisition of new texts and allow for linguistic analyses about the uses found with double-checking the text: in the context of course, and in the facsimile. It will also be possible to search the texts for keywords that constitute the Iconclass thesaurus, offering in this way a selection of topoi present in the texts, in their tables of contents (all transcribed) and in the illustrated elements.

5. In this way, the BVH is a "Renaissance" corpus that is still in the process of being built, a unique but double-headed corpus, one that is dissymmetric: there are obviously many more facsimiles than transcribed texts. Such a corpus is unified by the metadata that harmonize the catalogue of the work and that of the transcribed text thanks to appropriate descriptors (TEI headers, Dublin Core, OAI protocol) and by the search tools that are used on the plain text, but also thanks to the keywords associated with images and with texts, always offering as an anchor the facsimile and the reference to the original work.

Co-Reference: A New Method to Solve Old Problems

Øyvind Eide
University of Oslo
oyvind.eide@edd.uio.no

Introduction

In this abstract, I will describe co-referencing as a method for information integration. This method is based on building blocks that have been available in analogue as well as digital information systems for a long time. Co-reference is about connecting such building blocks in new ways in order to enable additional tools necessary for creating a functional Semantic Web for cultural heritage.

The co-reference concept will be explained and related to other tools for information integration. Some prototype implementations currently being developed for use in storing co-reference facts is being described, along with theoretical work towards the so-called *network of identity*. As a conclusion, I describe how we organise our work in this field as well as our plans for the future.

Definition

An example of a reference is the string "The table by the window," referring to a physical object. If another sting "The beautiful table" is referring to the same physical object, the two strings are said to co-refer.

A co-reference system will add relations directly between different references in texts. To do this, an identity relation has to be used. A possible identity relation for people could be:

> *A=B iff A and B are both human beings that have lived or are living, and the body of A is at any time located at the same place in the physical room as the body of person B.*

Once such an identity relation is established, the co-reference definition is simple:

> *If string x refer to A and string y refer to B, and A=B, then x and y co-refer.*

Recording Co-References

Computer software is developed to detect and disambiguate names and other referring strings in texts. They are quick and reliable, but at the cost of a high level of mistakes; not-detected items as well as false positives. Even when used on very structured data with additional information available to help the algorithm, such as a library catalogue example described in (Bennett 2006), the success rate in automatic data integration is far from acceptable: 70% hits with an error rate at almost 1%. Human beings, on the other hand, are more reliable. But they are also expensive.

There are two different ways of solving this quality-cost problem. One could use the computer to create candidate co-references and let human beings check the results, possibly as a federated job by amateurs. Or one could try to find ways to store results of work that people are already doing as formalised co-reference information.

As many people working in the culture heritage sector detect co-references as part of their job, we do not need to hire new people. What we need to do is to create computational and organisational systems so that they are able to store the facts they detect in a form usable for a co-reference system.

Information about the events in which co-references are asserted should also be stored. We will then be in a position to differentiate between statements made by computer programs and statements made by persons, as well as between different persons. This enables a higher degree of reproducibility than what is common in many areas of research made on the basis of cultural heritage. It also provides an addressee to approach if someone later suspects that a co-reference assertion may be wrong.

Similar approaches

Work related to co-reference detection is common in research. One example of tools used in such work today is thesauri. But instead of connecting the sources to one another, as one would do in the co-reference approach, each of the sources are linked to a record in the thesaurus. From such links to thesauri, co-references may be detected, but details about the event in which the source for the co-reference information was recorded is commonly not available.

Prosopographies are closely related to co-reference work, but has not traditionally published full sets of co-reference information about the names mentioned in the sources used. The factoid approach described by Bradley and Short, on the other hand, will provide information in a formalised way from propopographical work that is very close to co-reference data sets as they are described in this abstract: "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." (Bradley

2005, p. 8)

The concept of co-reference is used in corpus linguistics as well, in the sense that co-reference annotation is applied to text corpora (Day 2008). This is closely connected to anaphora resolution. One may say that anaphora resolution is a part of co-reference detection as an anaphor is just another referring string.

Level 1: Inside the Organisation

Co-reference management should be part of the overall information strategy in an institution. The curator, researcher or exhibition professional should be able to add information to their local system about external resources co-referring with internal resources.

Tracking down co-references have always been one of the practical tasks performed by researchers, conservators, librarians, and others processing information about real world objects described in texts. The results have sometimes been published, e.g. in indices or in footnotes. But most often it has been saved in the scholar's head, maybe also in his notes. We should do better. We should enable the creation of formalised data that may look like this:

ID	String	URL	Offset
Pers2345	John	http://foo.org/landregister.html#record2365	Position 5-8
Pers2345	Johannes	http://bar.org/smithetal/facs/page543.png	Rectangle (245,64,452,87)

The syntax of this example is not important. The information could e.g. be expressed in RDF. The point is that it is expressed in a machine readable way.

We are currently implementing a prototype co-reference system in which specific pieces of information in our databases can be connected to external resources. The external resources can be a URL, but can also be a reference to a specific place in a printed book. This system will also store information about who is responsible for asserting the co-reference, when it was done, and optional comments (Eide 2008). Further development is necessary to evaluate if this method will be usable in practical work.

Another implementation is the tagging tool created at FORTH in Greece as a diploma thesis by Kostas Pyloudis and Pasxalis Georgopoulos. The tagging tool is a web based application in which HTML web pages and photographs on the internet can be co-referenced and annotated by the user. The operation includes selecting a part of the document, so that part of an image, e.g. the head of one person, can be co-referred to a string in a HTML document, e.g. a name (Melesanakis 2008). This is an interesting attempt to enable anyone to take part in building up collections of co-reference information. Methods used by the Perseus Digital Library in their information integration work are also similar to the co-reference approach (Babeu 2007).

Level 2: Network of identity

After a while, many co-references have been stored in many institutions, some of them referring to the same individuals. All such co-reference collections should be connected and made available in a cross-institutional system.

The systems at level 1 described above will be sustainable, because the organisations gain from using them. The federated system will be the sum of all the two-way links from the level 1 systems. A model for such a system is described in (Meghini 2008). The article also describes a possible implementation. The system will be able to answer questions about the entities being co-referred. If it is connected to information systems storing information about e.g. events, it will be a very interesting tool for exploring large data sets based on cultural heritage information.

Conclusions

To integrate the wider cultural heritage field, it is necessary to connect persons, places, and objects described in and owned by museums to references in e.g. digital versions of printed books. It is not enough to connect classes of information, particular items will also have to be connected. Thus, co-reference tools will find their place among the other tools used in the development of the Semantic Web.

The CIDOC co-reference working group was established in 2007. We are currently working on the research and implementation described above. We are also developing prototype protocols for exchanging co-reference information between systems. We hope to set up a small integrated test system in 2009 in order to show how such systems may be implemented.

Literature

Babeu, A., Bamman, D., Crane, G., Kummer, R. and Weaver, G. (2007). "Named Entity Identification and Cyberinfrastructure." In: *Lecture Notes in Computer Science*, volume 4675.

Bennett, R., Hengel-Dittrich C., O'Neill, E.T. and Tillett

B.B. (2006). “USA VIAF (Virtual International Authority File): Linking Die Deutsche Bibliothek and Library of Congress Name Authority Files.” 72nd *IFLA General Conference*. URL: http://www.ifla.org/IV/ifla72/papers/123-Bennett-en.pdf (checked 2008-11-14).

Bradley, J. and Short, H. (2005). “Texts into Databases: The Evolving Field of New-style Prosopography.” P. 3-24 in: *Literary and Linguistic Computing*.

Day, D., Hitzeman, J., Wick, M., Crouch, K and Poesio, M (2008). “A Corpus for Cross-Document Co-reference.” P. 2996-2999 in: *Proceedings of the Sixth International Language Resources and Evaluation* (LREC’08).

Eide, Ø. (2008). “The Unit for Digital Documentation (EDD) system for storing coref information. A short overview of a system under development.” Paper presented at the meeting of the CIDOC Co-Reference Working Group, Athens. URL: http://cidoc.mediahost.org/co-reference-meetings-2008(en)(E1).xml (checked 2008-11-14)

Meghini, C., Doerr, M. and Spyratos N. (2008). “Managing co-reference knowledge for data integration.” P. 229-248 in: *Proceedings of EJC2008, the 18th European-Japanese Conference on Information Modelling and Knowledge Bases*.

Melesanakis, V. (2008) “Tagging-Tool.” Paper presented at the meeting of the CIDOC Co-Reference Working Group, Athens. URL: http://cidoc.mediahost.org/co-reference-meetings-2008(en)(E1).xml (checked 2008-11-14).

Cultural Capital in the Digital Era: Mapping the Success of Thomas Pynchon

Edward Finn
Stanford University
edfinn@stanford.edu

Overview

This paper will present some of my first work on a larger dissertation project: a new model for cultural capital in the digital era. In a time of rapidly evolving ecologies of reading and writing, I argue that the Internet affords us massive amounts of new data on previously invisible cultural transactions. New architectures for reviewing, discussing and sharing books blur the lines between readers, authors and critics, and these cultural structures capture thousands of conversations, mental connections and personal recommendations that previously went unrecorded. Through a close examination of the networks of cultural interchange surrounding the work of Thomas Pynchon, I hope to define a new, statistically informed conception of cultural capital and different modes of critical and commercial literary success.

Background

This paper takes as its starting point the idea that cultural capital can be defined or at least mapped online through exchanges at the intersection of literature as a field of art and the business of buying and selling books.[1] The advent of new media is transforming this landscape of literary production and consumption, making possible all sorts of new ecologies of reading and writing. At the same time, many transactions at the heart of digital cultural capital (i.e. customer reviews, shared libraries, “also-bought” lists, etc) are now captured in digital amber. Just as importantly, long-running statistical research into text-mining and content extraction has made effective tools available to map out semantic links between concepts and proper nouns in text archives. By combining data from digital ecologies and professional critical book reviews, I will compare the networks of cultural reception and critique that make for authorial fame. This combination of research methodologies promises new insights not only into the vast datasets of internet culture but also into the often-hidden cultural relationships that exist between various professional and popular archives, some of which reach back decades or centuries. There are exciting possibilities here for tracing the networks of influence and exchange that make up contemporary literary success and bringing some empirical rigor to often vague conceptions of cultural capital.

Proposal

As I work on the larger project of outlining a new, digital cultural capital, I propose a paper that explores some of these themes in the more limited arena of a particular contemporary author: Thomas Pynchon. Pynchon's pointed refusal to offer a public presence to the standard engines of publicity and authorial fame (he lives in more or less total secrecy, allegedly in Manhattan, and no clear photographs have been taken of him for decades) makes him an ideal candidate. Perhaps more than any other major contemporary author, he has relied on particular forms of cultural capital to earn literary fame, even as he critiques the social transformations wrought by capitalism through his work.

This paper will attempt to map the network of Pynchon's cultural capital and, through this exercise, make some larger claims about how cultural capital functions in the digital era. I will combine a broad reading of the discourses Pynchon engages in his novels with a study of the critical and popular reception of his work. If the statistical analyses I perform here cannot "prove" or "disprove" that reading, they will still inform it, just as the reading will inform my interpretations of Pynchon's cultural network maps. I see Pynchon as a bridge figure between the emerging discourses of post-War techno-supremacy, information theory, the new managerial class, and American consumerism. According to this position the allusive density and postmodern genre-bending of Pynchon's novels serve in effect to train readers in a form of interdisciplinary network reading. Dealing with such disparate and detailed threads of information, we all become paranoid connection-seekers, bridging different chasms and structures of knowledge across Pynchon's varied literary terrain. One question these results will necessarily address is to what extent Pynchon readers are more drawn to participation in his cultural networks online as compared to readers of other relatively similar contemporary authors (see methodology, below).

One of the most effectively reclusive authors on the contemporary scene, Pynchon has succeeded admirably in allowing public attention to focus solely on his work. I hypothesize that Pynchon's difficult, boundary-breaking books succeed in forging their own ideational networks, and that their author has both exploited and redefined contemporary cultural capital. In this way Pynchon places his readers in exactly the position they have assumed in new digital ecologies of reading and writing: the role of critic/operator, of consumer/creator, who grapples with unstable fictional ontologies from the vantage point of equally unstable critical ontologies.

Methodology

There is no comprehensive dataset or perfected analytical tool for exploring cultural capital, so this paper will construct its argument with several imperfect measures.

The argument will draw primarily on results from a new, wide-ranging dataset of reviews incorporating "highbrow" cultural review publications like *The New York Times*, academic journals, popular media and customer reviews from blogs and websites such as Amazon.com. Using a combination of APIs, Perl scripts and commercial databases, I am assembling a targeted dataset of professional and popular reviews and recommendations from the book review archives of major newspapers, magazines, and the websites Amazon and LibraryThing, with more academic, popular and commercial archives to follow should time permit.[2] Using named entity recognition (NER) and part of speech tagging (PoS), I will construct social and conceptual network maps based on this corpus to explore a) how Pynchon was introduced or integrated into the existing cultural firmament, b) who reviewed and/or was mentioned in connection with Pynchon and how these networks evolve, c) what insights I can draw from resonances and differences in the conceptual maps drawn from different categories of cultural reaction (e.g. professional reviews vs. reader reviews). This analysis will adapt an available open-source NER and PoS text-mining tool (most likely Carnegie Mellon's *AutoMap* or the University of Massachusetts' *Proximity*) to construct the network maps.

These maps, while hopefully revelatory, will not offer much in the way of quantitative measurement of social capital or literary success. To address the problem of quantifying cultural capital, I will be creating control groups of books for each Pynchon work discussed. These groups of 5-10 novels will be selected based on their cultural proximity to the relevant Pynchon work (a messy term, I concede): published within a year of the Pynchon work; written by an author with similar style, interests and/or who might be reasonably said to compete with Pynchon; greeted with similar levels of highbrow acclaim, and greeted (to the extent this can be determined) with similar levels of commercial success. It is my hope that the potential liabilities of any particular judgment or selection will be ameliorated in aggregate, and that the control groups will still provide a reasonable baseline for studying how Pynchon's work has fared over time.

Using these baselines, Pynchon's work will also be considered along several more quantitative metrics. I will trace the historical reception of his work through measures of annual MLA citations and popular press citations (the latter will most likely use a small index of publications—newspapers and magazines with consistent national presence over the past 50 years). Since sales figures are notoriously difficult to come by, I will attempt to chart the current popularity and lasting success of Pynchon's books by looking at library presence (through meta-catalogs like Worldcat.org), used books available for purchase online, and presence on library sharing sites like LibraryThing. Of course these popularity statistics will be judged comparatively against their normative control groups. Finally, I should note that while the control groups will allow some metrics to be tracked over time (citations, for example), used book availability, library presence and some other factors will necessarily be limited to a snapshot of current conditions.

Conclusion

By exploring contemporary networks of cultural capital related to the work of Thomas Pynchon, I hope to develop tools and statistical methodologies that will be adaptable to other authors in my own project and to other scholars for their own work. I am aware of limited applications of social and conceptual network mapping in literary study, but in extensive searches I have yet to find an easily adaptable tool that non-technical literary scholars might use to explore a new corpus. I hope to close this gap at least partially as I make my own arguments about the evolving nature of digital cultural capital.

Notes

The notion of cultural capital discussed here is based most directly on the work of John Guillory in *Cultural Capital: The Problem of Literary Canon Formation*, and through him Pierre Bourdieu. The term is a fraught one and a full definition is beyond the scope of this paper. Instead, I hope to lay the groundwork for a new, digital understanding of cultural networks and influence as they relate to literary production in the larger capitalist system, using cultural capital as a bracketed general term.

To briefly explain my choice of websites: Amazon is an obvious choice because of its active reviewing community and relatively long-running archive, with over a decade of reviews available. *Publishers Weekly* declared LibraryThing the most popular of the social reading network sites, and it also encourages an active group of user-reviewers. A more extensive study will, with luck, involve other major book-reviewing sites online.

Selected Bibliography

Bourdieu, P. (1984). *Distinction: A Social Critique of the Judgment of Taste*. Trans. Richard Nice. Cambridge, MA: Harvard University Press.

Bourdieu, P. (1993). *The Field of Cultural Production: Essays on Art and Literature*. Ed. Randal Johnson. New York: Columbia University Press.

English, J. (2005). *The Economy of Prestige*. Cambridge, MA: Harvard University Press.

Guillory, J. (1993). *Cultural Capital: The Problem of Literary Canon Formation*. Chicago: University of Chicago Press.

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

Radway, J. (1997). *A Feeling for Books: The Book-of-the-Month Club, Literary Taste, and Middle-Class Desire*. Chapel Hill: University of North Carolina Press.

The Hybrid Future of the University Press

Kathleen Fitzpatrick
Pomona College
kfitzpatrick@pomona.edu

Numerous arguments have been put forward in recent years about the causes and effects of the crisis in scholarly publishing, as have numerous more suggestions about ways to ameliorate the situation (one might see, just to name a few, Waters (2004), Alonso et al (2003), and Greenblatt (2002)), but few of these accounts seem to get at the heart of what is, admittedly, a very thorny problem in academic publishing today: an utterly insupportable business model. This paper, which forms a small part of a book-length project exploring the social and institutional changes required to make digital scholarly publishing a reality, will not argue for ways of creating supports for the existing system (whether through subventions, book-buying funds, or other means of funding production or increasing consumption). And though it follows the work of authors including Willinsky (2006), Borgman (2007), and Hall (2008) in arguing that the future of scholarship must be digital, this paper will not argue that such a turn to digital publication can in and of itself rescue scholarly publishing from its financial crisis. Instead, this paper will argue, scholarly presses must consider a far more radical shift in their business models, in which they cease thinking of themselves as providers of products for sale, and instead understand the publisher as a provider of services that facilitate scholars' interactions with texts, and with one another through those texts.

Clay Shirky argued as long ago as 1997 that an internet-based business model focused on the sale of content was destined to fail. The shift from content to services, however, might best be understood within the model of the "hybrid" economy described by Lawrence Lessig (2008). The hybrid is neither a wholly commercial nor a wholly gift-based economy, but rather one that creates value for users by offering services they desire, thereby encouraging them to contribute their labor to the enterprise. Lessig explores the models established by several successful hybrid businesses, including Flickr, Slashdot, Craigslist, and others, suggesting that contemporary content providers (like the music industry) who have the sense that their bottom lines are being undermined by network-based file sharing would do well to consider the ways that their business models might change in order to take advantage of peer-to-peer networks rather than attempting to legislate or sue them out of existence.

The concerns of university presses about the digital future are slightly different, but these presses nonetheless might take advantage of these same principles. The relationship, after all, of authors to the university press is already based at least in part on the culture of the gift; few academic authors earn much directly from their published texts, instead benefitting from the jobs and speaking engagements that their publications produce. And in the digital age, as Bob Stein (2008) has explored, presses will need to think less about selling the content of texts they publish, and instead focus on the services that they can provide to authors, in the development of their texts, and to readers, in providing means of interacting with the texts, and with one another around the texts.

Implied in this turn from products to services, however, is a large-scale shift in the relationship of the university press to its institution, as explored in Brown et al (2007). Despite the fact that most U.S. based university presses arose out of the desire of the institution to publicize the work of its faculty, most presses today operate on a list-based model, primarily publishing the work of scholars from other institutions, and focusing on a select number of fields. The result is precisely the untenable business model faced by presses, which are expected to operate as businesses rather than service organizations (even where they are subsidized, if only minimally, by their institutions). Changing the focus of the press from selling the products of scholarly research to facilitating the processes of that research will also require the academic institutions that house presses to recognize that a press that functions as a service organization within the university cannot simultaneously serve as a revenue center. I will argue that the survival of the university press in the current economic and technological climate will require that presses return to their earlier, service relationship to authors within their own institutions, in order to more firmly cement their position within the heart of the university's overall mission.

In these two respects — in turning from selling the products of scholarship to facilitating the networked means through which scholarship is done, and in shifting its focus to serving the needs of their institutions, presses might learn from libraries — and might, as Crow (2009) argues, benefit from becoming more strongly allied with libraries, as one serves the needs of the institution by gathering published material from around the world for its users, and the other serves those needs by distributing locally-produced texts to users around the world. As with libraries, however, this new position of the press within the university's overall mission will require that institutions fund their presses as part of their infrastructure, rather than understanding the press as a revenue center — and, not incidentally, it will also require that institutions without presses establish them in order to remain competitive.

This presentation will thus explore not the new technologies that will rescue the university press, but rather the new business model that those technologies will require the press to develop, arguing that a focus on services rather than products, and a new relationship to the university's core mission, will enable the press, as a nexus for new modes of scholarly communication, to thrive into the future.

References

Alonso, C. J., Davidson, C. N., Unsworth, J., & Withey, L. (2003). *Crises and Opportunities: The Futures of Scholarly Publishing*. New York: American Council of Learned Societies.

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

Brown, L., Griffiths, R., & Rascoff, M. (2007). University Publishing in a Digital Age. Ithaka. First published 23 July 2007, accessed 17 June 2008, http://www.ithaka.org/strategic-services/university-publishing.

Crow, R. (2009). Campus-Based Publishing Partnerships: A Guide to Critical Issues. SPARC. First published January 2009, accessed 8 February 2009, http://www.arl.org/sparc/partnering/guide/.

Greenblatt, S. (2002). A Special Letter from Stephen Greenblatt. Modern Languages Association. First published 28 May 2002, accessed 28 February 2009, http://www.mla.org/scholarly_pub.

Hall, G. (2008). *Digitize This Book!: The Politics of New Media, or Why We Need Open Access Now*. Minneapolis: University of Minnesota Press.

Lessig, L. (2008). *Remix: Making Art and Commerce Thrive in the Hybrid Economy*. New York: Penguin.

Shirky, C. (1997). Help, the Price of Information Has Fallen and It Can't Get Up. *Clay Shirky's Writings About the Internet*. First published April 1997, accessed 28 February 2009, http://www.shirky.com/writings/information_price.html.

Stein, B. (2008). A Unified Field Theory of Publishing in the Networked Era. *if:book*. First published 4 September 2008, accessed 1 November 2008, http://www.futureofthebook.org/blog/archives/2008/09/a_unified_field_theory_of_publ_1.html.

Waters, L. (2004). *Enemies of Promise: Publishing, Perishing, and the Eclipse of Scholarship*. Chicago: Prickly Paradigm Press.

Willinsky, J. (2006). *The Access Principle: the Case for Open Access to Research and Scholarship*. Cambridge, Mass: MIT Press.

Dissent and Collaboration

Julia Flanders
Brown University
Julia_Flanders@brown.edu

Collaboration—literally a shared work—is typically understood to rest upon a form of agreement: about shared goals, common projects, standards of practice. As Aldo de Moor observes, modern-day collaborative practices tend to emerge from self-forming teams rather than being organized from above, and as a result this kind of agreement constitutes an essential foundation on which to proceed. He goes on to note that the crucial element to successful digital collaboration has a great deal to do with the way norms are developed and adjudicated:

> The members of virtual professional communities, just like their peers in more traditional communities, are guided in their work by social norms. These norms guide both the operational activities of the network and the specification processes in which the network is defined. However, as these networks are egalitarian by nature, the norms cannot be imposed from above, but are to emerge from the community as a whole. Thus, the user-driven specification process needs to be legitimate, in the sense that specification changes are not only meaningful but also acceptable to all community members. A specification change is acceptable if and only if the users for whom the particular change is relevant have been adequately involved in the specification process (Aldo de Moor, "Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems", summary of PhD thesis, http://www.communitysense.nl/phd/thes_summeng.html).

Relevant norms in the digital humanities context are quite wide-ranging: they include social norms, disciplinary norms, and technical norms. But what I would like to focus on for purposes of this paper is a specific kind of case in which disciplinary and technical norms overlap: the arena of standards for digital representation of research materials, and in particular the domain of scholarly markup languages. Text encoding as practiced in the digital humanities world sits at the juncture of humanities scholarship—textually nuanced, exploratory, and introspective—and digital technology, with its emphasis on formalism and upward scalability. As a result its norms carry a double weight: they must achieve some kind of technically actionable uniformity, but they must also express useful scholarly concepts and differentiations.

Encoding standards such as the Text Encoding Initiative (TEI, http://www.tei-c.org/) are thus foundationally collaborative technologies: they presume the need and the desire to make individual insight widely communicable in a form that permits its extension, critique, and reuse. But the mechanisms for achieving this result in practical terms are complex and require thoughtful balancing of the requirements, respectively, of the individual and the community. As de Moor observes in the material quoted above, legitimate norms arise from a process of community assent, but how is that assent best managed and expressed? And in the universe of humanities disciplines, where "the community" consists of multiple communities with shifting boundaries, how are norms constituted in a way that still permits intellectual growth? Even more importantly, given the critical role that dissent and debate play within the humanities research context, how can these be expressed? Can we imagine a role for dissent within a functioning technical standard, without vitiating its power to support collaboration?

Over the past 20 years, the research of the TEI and its user community has been centered on developing mechanisms that address the problem of norm-setting, in both the social and the technical sphere, in a way that I will argue is specifically designed to accommodate dissent in a way that actively facilitates collaboration. This paper will critically examine the TEI's framing motivations and the specific mechanisms—intellectual, social, and above all technical—through which they have been realized during the course of the TEI's development. In particular I will consider the practice of schema customization, through which the TEI manages both the representation of the TEI language as a standard and the processes of dissent and expansion through which it is modified by its users. The central components of the TEI customization process express, in effect, the relationship between the individual and the community. The TEI source or ODD file represents the entire landscape of the TEI in potential terms, and can be used to generate a maximally capacious schema that contains all possible TEI elements and structures. The ODD customization file represents the world of an individual user or project: the set of choices and limitations or extensions through which the individual adapts the TEI schema to local usage. From these two files, with appropriate processing, one can generate a schema that expresses the TEI landscape as viewed through the lens of the individual application. The ODD customization file, then, expresses the distance between the individual and the community: both the agreement the community has established concerning meaning, and also (perhaps more importantly) the degree by which the individual dissents from the standard, while still acknowledging its centrality for the community as a whole.

Dissent is nothing new in the humanities; what is distinctive about this mechanism is that it formalizes dissent and allows its vector to be traversed in two directions. The same path that leads away from the unmodified TEI standard towards the individual application (from generality to specificity) can also be followed back to the center again. This traversal can be effected both by human beings and by computer processes. Information concerning what has been changed and why can be expressed in human-readable form and may serve as a valuable support in understanding the methodological choices that underlie a project's encoding practice. Similarly, the ODD customization file can serve as the basis for automated analysis that could, for example, identify all projects from a large set that use the same set of TEI modules or remove the same set of elements; generate a list representing the greatest common set of values for a given attribute across a group of projects and also identify the values that are unique to each project; identify the range of new elements created by each project and their TEI equivalents. Taken as a whole, the customizable approach taken by the TEI permits the standard to function (both socially and technically) as an agreement at many levels—on the intention to treat data as a sharable and preservable resource, on the value of shared data standards, on the descriptive utility of this particular approach to modelling humanities texts, and on the impossibility of creating a single descriptive model that will satisfy all needs.

In an important sense, this customization mechanism encapsulates the central challenge of collaborative work, and even of language itself: that of how to balance the urge toward individual expressiveness with the mandates of public comprehensibility, the desire for individual freedom of agency against the need for group action. This paper will explore several specific examples drawn from real TEI projects to demonstrate the different kinds of managed dissent that can be expressed using the customization mechanism, including the following:

- the use of shared generic structures (<div>, <seg>, etc.) coupled with customized constraint of attribute values to express variations on common structures

- the use of "syntactic sugar": the creation of project-specific elements that are explicitly equivalent to standard TEI elements

- the use of the TEI class system to express how a project-specific vocabulary fits into the TEI's structural model

The important question emerging from these issues is what this kind of discursive agreement and dissent actually achieves in the collaborative domain. Several points should be noted and will be explored in more detail in the finished paper. First, agreement of course permits communication, but properly formalized dissent is an equally important dimension of communication. Language (including formal encoding language) expresses a view of the world, and it is essential to be able to disagree about that view, or to express a different view. But that disagreement must be explicit rather than covert: otherwise tag abuse is the result. A second, complementary question is how we can distinguish between meaningful dissent (essential points of disciplinary difference) and meaningless dissent (based on laziness or other social failures). Are the mechanisms for making this distinction solely social and human, or are there automatable mechanisms conceivable as well? Finally, what kinds of collaboration are possible where formalized dissent and disagreement, rather than complete agreement, are the result of using a standard? I will argue that for the digital humanities community, effective management of disagreement—rather than simply the production of agreement—is the most important role a standard like the TEI can play.

Science Fiction in the Lives of Scientists and Engineers

Kenneth R. Fleischmann
University of Maryland, College Park
kfleisch@umd.edu

Thomas Clay Templeton
NASA Goddard Space Flight Center
thomas.c.templeton@nasa.gov

Introduction

This study explores the influence of science fiction read in childhood on the career choices and research trajectories of scientists and engineers. Mosco (2004) argues that science fiction, as modern-day myths, significantly impacts how people think about and view the world. Scholars of science fiction argue that science fiction is an influential element of contemporary American society (Disch, 1998; Malmgren, 1991; Stableford, 1987). Specifically, one interesting facet of this issue is the role of science fiction in influencing public understanding of and attitudes toward science (Chaloner, 1998; Claessens, 2004). Indeed, one additional significant societal implication of science fiction is the impact it can have in structuring and changing individuals' lives (Bacon-Smith, 2000; Jenkins, 1992). Science and literature studies examine the relationship between technoscience and science fiction on a conceptual level and thus serve as an important inspiration for this study (Dery, 1996; Doyle, 1997; Haraway, 2004; Hayles, 1999), but have not undertaken empirical research to examine the impact of science fiction on the lives of scientists and engineers. This focus on the role of science fiction in the lives of scientists and engineers is part of the larger theme of the book in the life of the reader, or how literature affects readers' everyday lives (Andringa and Scheier, 2004; Kaestle, 1991; Pawley, 2002; Polhemus and Henkle, 2005; Wiegand, 1998). This study seeks to determine if and how science fiction affects technoscientific research.

Methods

Data collection included fourteen semi-structured interviews with scientists and engineers as the NASA Goddard Space Flight Center. Interviews were conducted using an oral history approach (Baumgartener and Payr, 1995; Lyons and Taksa, 1982; Radway, 1991; Ritchie, 2003; Sommer and Quinlan, 2002; Thompson, 2000; Yow, 2005). Interviews were recorded, transcribed, and analyzed using grounded theory (Strauss and Corbin, 1998). Specifically, the researchers first independently coded interviews line-by-line to identify specific factors that influenced the career choices of scientists and engineers. The researchers then compared their two lists of factors and used them to build a list of clusters that encompassed multiple factors.

Results: Factors Influencing the Career Choices of Scientists and Engineers

This research identified ten predominant clusters of factors that NASA Goddard scientists and engineers described as influencing their career choices: the space program, science fiction, informal education, formal education, the job market, serendipity, personal attributes, values, family and friends, and the physical environment. Each of these clusters was identified in multiple interviews, and most of the factors within each cluster were also found in multiple interviews. The complex interconnections among these diverse factors are depicted in Figure 1.

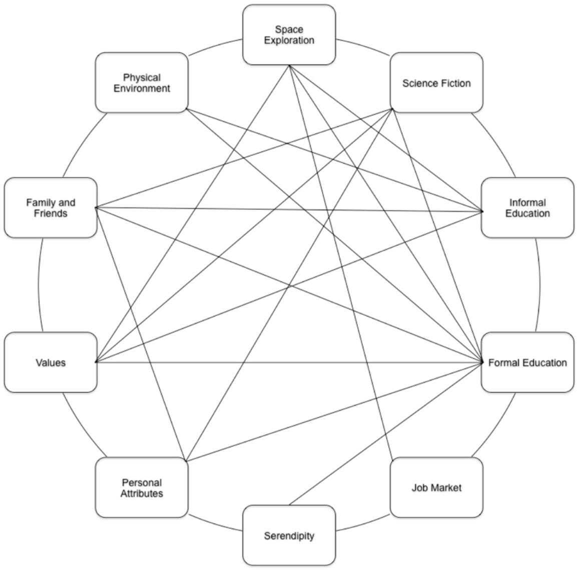


Figure 1: Clusters of Factors that Influence the Career Choices of Scientists and Engineers

The space program included factors such as the space race, from the launch of Sputnik to the first successful Apollo Moon landing, as well as more recent events such as the space shuttle program. Science fiction included both 'hard' science fiction by authors such as Asimov, Clarke, and Heinlein and popular forms of science fiction such as Star Wars and Star Trek. Informal education included factors such as museums, kits, and books. Formal education included factors such as K-12 and undergraduate science courses. The job market included factors such as the availability of jobs in different areas at different times. Serendipity included factors such as whom the interviewees knew and how one life event led to another. Personal attributes included factors such as gender, interest in different levels of abstraction, and financial considerations. Values included factors such as curiosity and altruism. Family and friends included factors such as the career choices and influence of parents, siblings, grandparents, and friends. The physical environment included factors such as fascination with the night sky and with nature parks.

Results: Themes for the Influence of Science Fiction on Technoscience

The findings of this study demonstrate that science fiction can influence the career choices of scientists and engineers. Scientists and engineers explained that science fiction, "kept me interested in science and astronomy;" "may be a lot of people's first experience with real science;" "may keep you inspired;" and "continues to spark people's imaginations." A total of eight major themes for the influence of science fiction on technoscience were identified: worldview/perspective, ideas/insights/inventions, inspiration, initial exposure, shared culture/shared meaning/socialization, reinforcement, excitement, and curiosity.

Limitations

As in the case of all interview research, data may be influenced by self-report bias. Further, inquiry about childhood influences on professional development are highly speculative, as it is often not possible to discern among the many factors that influence decision-making. Further, with such a small sample, it is not possible to determine the likelihood of influence by science fiction and other factors; rather, it is only possible to report on the self-reported experiences of this small sample of scientists and engineers.

Future Research Directions

One planned future direction is to conduct a broad survey to test hypotheses developed as a result of the interviews. This planned study would include a larger sample of scientists and engineers from a broader range of organizations and fields. Another possible future direction is to develop educational interventions that use science fiction to interest undergraduate students in science and engineering, building on earlier work in this area (Berne and Schummer, 2005; Raham, 2004). This study provides empirical evidence that science fiction influences the career choices and research trajectories of scientists and engineers, and points toward ways to use such educational programs to increase the size and diversity of the science and engineering workforce.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant SES-0731819. The authors would also like to thank the fourteen scientists and engineers who participated in this study.

References

Andringa, E., and Scheier, M. (2004). How Literature Enters life: An Introduction. *Poetics Today*, 25: 161-169.

Bacon-Smith, C. (2000). *Science Fiction Culture*. Philadelphia: University of Pennsylvania Press.

Baumgartner, P., and Payr, S. (1995). *Speaking Minds*. Princeton, NJ: Princeton University Press.

Berne, R. W., and Schummer, J. (2005). Teaching Societal and Ethical Implications of Nanotechnology to Engineering Students Through Science Fiction. *Bulletin of Science, Technology, and Society*, 25: 459-468.

Chaloner, P. A. (1998). Chemistry in TV science fiction: *Star Trek* and *Dr. Who*. In J. H. Stocker (Ed.), *Chemistry and Science Fiction*. Washington, DC: American Chemical Society.

Claessens, M., ed. (2004). Science Fiction: Intuition and Fantasy. *RTDinfo: Magazine for European Research* Special Issue, (March), 21-23.

Dery, M. (1996). *Escape Velocity: Cyberculture at the End of the Century*. New York: Grove Press.

Disch, T. M. (1998). *The Dreams Our Stuff Is Made Of: How Science Fiction Conquered the World*. New York: The Free Press.

Doyle, R. (1997). *On Beyond Living: Rhetorical Transformations of the Life Sciences*. Stanford, CA: Stanford University Press.

Haraway, D. J. (2004). *The Haraway Reader*. New York: Routledge.

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

Jenkins, H. (1992). *Textual Poachers: Television Fans and Participatory Culture*. New York: Routledge.

Kaestle, C. F. (1991). The History of Readers. In Kaestle, C. F.; Damon-Moore, H.; Stedman, L. C.; Tinsley, K.; and Trollinger, Jr., W.V. (eds), *Literacy in the United*

States: Readers and Reading Since 1880. New Haven, CT: Yale University Press.

Lyons, M., and Taksa, L. (1992). *Australian Readers Remember: An Oral History of Reading 1890-1930*. New York: Oxford University Press.

Malmgren, C.D. (1991). *Worlds Apart: Narratology of Science Fiction*. Bloomington, IN: Indiana University Press.

Mosco, V. (2004). *The Digital Sublime: Myth, Power, and Cyberspace*. Cambridge, MA: The MIT Press.

Pawley, C. (2002). Seeking 'Significance': Actual Readers, Specific Reading Communities. *Book History*, 5: 143-160.

Polhemus, R. M., and Henkle, R. B., eds. (2005). *Critical Reconstructions: The Relationship of Fiction and Life*. Stanford, CA: Stanford University Press.

Radway, J. A. (1984). *Reading the Romance: Women, Patriarchy, and Popular Literature*. Chapel Hill, NC: University of North Carolina Press.

Raham, R. G. (2004). *Teaching Science Fact with Science Fiction*. Portsmouth, NH: Heinemann.

Ritchie, D. A. (2003). *Doing Oral History, A Practical Guide, Second Edition*. Oxford, UK: Oxford University Press.

Sommer, B. W., and Quinlan, M. K. (2002). *The Oral History Manual*. Walnut Creek, CA: AltaMira Press.

Stableford, B. M. (1987). *The Sociology of Science Fiction*. San Bernadino, CA: The Borgo Press.

Strauss, A. and Corbin, J. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, Second Edition*. Thousand Oaks, CA: Sage.

Thompson, P. R. (2000). *The Voice of the Past: Oral History, Third Edition*. New York: Oxford University Press.

Wiegand, W. A. (1998). Theoretical Foundations for Analyzing Print Culture as Agency and Practice in a Diverse Modern America. In J. P. Dansky and W. A. Wiegand (eds), *Print Culture in a Diverse America*. Urbana, IL: University of Illinois Press.

Yow, V. R. (2005). *Recording Oral History: A Guide for the Humanities and Social Sciences, Second Edition*. Walnut Creek, CA: AltaMira Press.

Creating a Composite Cultural Heritage Artifact – the Digital Object

Fenella G. France
Library of Congress
frfr@loc.gov

Eric F. Hansen
Library of Congress
ehan@loc.gov

Michael B. Toth
R. B. Toth Associates
mbt.rbtoth@gmail.com

Advanced digital spectral imaging is now becoming a key interdisciplinary analytical tool for use in the examination, assessment and understanding of cultural heritage preservation and interpretation in the humanities. This involves the integration of advanced technologies, work processes and a range of skills from various fields. This includes experts in materials science and information technology with social scientists through the development of methodologies to ensure preservation, distribution and access to all aspects of cultural heritage. The critical component of developing this research area is to develop collaborations between researchers in multiple disciplines. The Library of Congress is developing and refining this tool for preservation of significant cultural heritage artifacts, including the Waldseemüller 1507 World Map, the first and second drafts of the Gettysburg Address, and the L'Enfant (1791) plan of Washington D.C. Under the direction of the senior preservation scientist, Dr. Fenella France, a team of preservation scientists and imaging specialists developed imaging and data management capabilities to ensure preservation, distribution and access to information on these and other cultural heritage objects. This information proved critical in supporting collaborations between researchers in multiple disciplines, enhancing the ability to understand and mitigate the potential and inevitable loss of information due to the natural degradation of materials.

Hyperspectral imaging involves the capture of a range of specific wavelengths in the visible and non-visible spectrums. To access this wealth of image data, three key aspects of digital data capture and management must be addressed:

- The image acquisition process and the required adaptation for high quality images of large documents and manuscripts
- Processing of digital images integrated with specific interpretation for preservation and scholars
- Implementation of effective metadata standards and data management systems

This paper will primarily address the integration of digital image acquisition and processing to develop a non-destructive analytical technique that allows the characterization of inks, colorants, treatments, substrates, deterioration, and lost information – all critical elements for preservation.

Imaging research at the Library of Congress includes the collection of sequential narrow wavelength bands of images from the ultraviolet, through the visible spectrum to the infrared (approximately 300nm – 1100nm). This hyperspectral imaging collects contiguous wavelengths, detecting variances in responses from materials in the artifact at any wavelength or combination of wavelengths. The team of preservation and imaging scientists working with a MegaVision 39 Megapixel monochrome camera and Equipoise LED EurekaLights image the items of cultural heritage – documents, leaves of manuscripts, daguerreotypes, in each spectral band. This yields a large data cube of digital images and metadata for each artifact. The resulting collected images are digitally combined with or subtracted from each other to form the processed image or "digital object" that provides greater analysis and interpretation of the original real cultural artifact. These processed images contain a wealth of information, but it is the significant levels of multidisciplinary interpretation required to process and analyze the data collected, that forms the new digital object.

The Library of Congress is building on the past two decades of characterization of ancient texts and documents using advanced digital imaging techniques. It is developing these research methods to advance studies of other media and cultural heritage items, as well as new techniques for specific preservation requirements. The ongoing research and development highlights the utility of transferring advanced imaging techniques developed for defense and astronomical studies to the preservation of significant cultural heritage artifacts. Specific studies that have led to the current developments in advanced digital imaging include the Dead Sea Scrolls, the Khaboris Codex, Archimedes Palimpsest, and the Oxyrhynchus Papyri. Each of these studies has revealed some of the issues and challenges involved with applying advanced imaging and processing techniques to researching cultural heritage documents and artifacts. For example, complex image processing previously used to recover images from astronomical telescopes, has been adapted to reveal what the original woodblock that was used in the printing of the Waldseemüller map probably resembled. Integrating an analysis of the techniques used by a woodcut maker in the 1500s informs understanding of materials and choices made even though the original woodcut no longer exists. All of this research underscores the need to integrate social and material sciences with the digital object to better understand issues of technological and other choices in the creation of the original object. Styles and the delicate detail attained were compared with copper engraving by researchers in the 20th century still attempting to understand the innate technical skills. Layering the information to show later additions by cartographers and printers adds more to the complexity and potential utilization of the digital object. The ability to process this information needed to collect useful data associated with substrates and media without physical sampling is critical to the assessment and preservation of many international items of cultural heritage. In addition, analysis for transcription and translation of deteriorated ancient texts requires collaboration between conservation and scholars for effective translation. These aspects will be outlined in more detail in the paper.

While advances in image acquisition processes are critical, it is the post-capture processing that creates the new digital object of interest in this discussion. This processing and interpretation of acquired images is integral to the creation of the information necessary for the preservation, analysis and assessment of these cultural artifacts through digital spectral imaging. This digital image processing has evolved from simply choosing the best spectral band and collecting images in that band (mono-spectral), through principle component analysis of specific wavelength combinations and pseudo-color processing, to the current advanced algorithmic image processing utilizing the combination of registered images in various non-visible and visible spectrum regions.

As noted previously, hyperspectral imaging creates large volumes of data for processing, analysis and interpretation and requires understanding of these new digital objects and the levels of information contained within. Data and metadata management is imperative to integrating digital imaging and processing capabilities for studies of cultural artifacts in the humanities. Creating the digital object requires effective systems and information management to ensure that the large amounts of digital data generated can be readily acquired, stored, archived, accessed, processed and linked to other data.

The creation of a new "digital object" that combines the real cultural artifact and the processed digital files and information has established a new category of cultural artifact in the humanities. The resulting composite digital object allows insights into methods of construction, the impact of society and technology and scholarly information on how this informs researchers and our understanding of previous societies. This can be used to address questions such as:

- Why did artists and artisans develop and employ those specific techniques that were used to create the real artifact?
- How were these restricted or informed by available tools, cultural norms, religious or moral beliefs?
- Did the creator demonstrate a break from tradition that was only now being revealed through the availability of advanced spectral information?

Integration of information from the original object and the digital imaging is a critical component in the organization and access to the digital object. Developed from GIS, the concept of "scriptospatial" or "image-spatial" tagging of key points on the images is the equivalent of a global positioning system for documents and objects. The linkage of information from other scientific analyses, or scholarly interpretations of revealed details revealed by the advanced digital image processing is an important element in the generation of an integrated digital object that allows the interpretation and assimilation of a range of data that is at times in different data sets.

The critical breakthrough in the adoption of hyperspectral digital imaging was the maturity its development – resolution, integrated conservation safe lighting and management – to concentrate on and answer these questions, while also addressing the issues involved for this technique to become accepted as a true non-invasive non-destructive analytical tool with no risk to fragile historic artifacts in the field of cultural conservation. Hyperspectral imaging has the capacity to reveal information, data and details not visible or accessible from the historic artifact itself due to deterioration. This highlights the necessary integration between materials science, digital spectral imaging and the social implications of retrieving lost information. This generation of new information includes that created from advanced digital image processing of the large volume of acquired data.

In conclusion, providing useful information to support conservation research and scholarly studies requires the effective integration and application of new technologies, work processes and technical skills to the field. The use of multi- and now hyperspectral imaging as an effective conservation tool has allowed the development of non-destructive analytical tools that allow the safe analysis and examination of texts and documents. This requires the integrtion of a range of associated activities and processes: imaging artifacts in a range of spectral bands, capturing important metadata about the digital records, storing the digital data and associated metadata, processing the images and data to yield useful information, and making the information available for researchers in the humanities, conservation professionals and the public. Future accomplishments in the creation of a new digital object will be dependent on economic development of integrated image information systems, continued advances in image technology, and the effective integration of data access, storage, management and interpretation. This requires continuing innovation and collaborations of imaging and preservation scientists, information technology professionals, conservators and researchers in the humanities.

References

Casini, A, et al. Image Spectroscopy Mapping Technique for Noninvasive Analysis of Paintings, Studies in Conservation 44 (1999) 39-48

Easton, R.L. Jr., Knox. K, Christens-Barry, W.A. Multispectral imaging of the Archimedes palimpsest. In: Proceedings of 32nd Applied Imagery Pattern Recognition Workshop (IEEE-AIPR'03) (2003) 111–116

France, F.G. Managing digital image repositories as key tools in the preservation of cultural objects, Imaging Science and Technology Conference, Arlington, VA, 2007

France, F.G. and Toth, M.B. Developing cultural heritage preservation databases based on Dublin Core data elements, Dublin Core Conference, Manzanillo, Mexico, October 2006

Grenacher, F. The Woodcut Map: A form-cutter of maps wanders through Europe in the first quarter of the sixteenth century, Imago Mundi, Vol 24 (1970) 31-41

Knox, K. "Enhancement of overwritten text in the Archimedes Palimpsest," in *Computer Image Analysis in the Study of Art, San Jose,* California, Proc. SPIE, vol. 6810 (2007)

Knox, K.T. et al., "Image Restoration of Damaged or Erased Manuscripts", European Signal Processing Conference, Lausanne (2008)

Knox, K. and Easton, R.L. Jr.: Recovery of lost writings on historical manuscripts with ultraviolet illumination. In: Fifth International Symposium on Multispectral Colour Science (Part of PICS 2003 Conference), Rochester, NY (2003) 301–306

Plaza, A, et al., Recent Advances in Techniques for Hyperspectral Image Processing, Remote Sensing of Environment, Elsevier Science, July 2007

Reedy, C. L. and Reedy, T. J. Relating visual and technological style in Tibetan sculpture analysis, World Archaeology 25(3) (1994) 304-320

Toth, M.B. "Management of Digital Archives for Integrated Web Access to Scientific and Cultural Information", Society for Imaging Science and Technology, Archiving Conference, Arlington, Virginia, May 21-24 (2007)

Toth, M.B., Emery, D. "Encoding Archimedes Work with TEI to Complete a 10-Year Program", Text Encoding Initiative Annual Members Meeting, London (2008)

Walvoord, D, Easton, R. Jr., "Digital Transcription of the Archimedes Palimpsest", *IEEE Signal Processing Magazine* p. 100-104, July (2008)

Digital History Across the Curriculum

Amanda L. French
New York University
amanda.french@nyu.edu

Peter J. Wosh
New York University
pw1@nyu.edu

If digital humanities is an isolated 'enclave' that remains largely isolated from humanities disciplines, as Martha Nell Smith averred at the 2008 *Digital Humanities and the Disciplines* conference, perhaps one reason is that the residents of that enclave have not often tried to initiate broad curricular reform. Certainly, researchers who are part of the digital humanities community teach courses that use the writings of digital humanists, courses that use the texts and images and audio and video digitized with funds granted to digital humanists, courses that use the web sites and databases and software developed by digital humanists, courses that use the theories and methods and language of the digital humanities, but such courses seem to remain the specialized offerings of specialists. Unlike women's studies, for instance, which at the very least increased the number of objects for humanities inquiry by an order of magnitude, the research-oriented field of digital humanities has not changed most humanities syllabi.

This paper will tell the tale of one graduate program's programmatic attempt to incorporate the intellectual and practical insights of the digital humanities throughout its curriculum. In 2008, the National Historical Publications and Records Commission awarded a grant to the Archives and Public History graduate program in the History department at New York University for the purpose of creating 'a model curriculum that fully engages new media' with 'a completely integrated and coherent approach to digital and electronic records issues.' (Wosh, 2007) To create this model curriculum, the program hired a Digital Curriculum Specialist to take on four well-defined tasks: first, to revise the syllabi for three key courses and to work closely with faculty and graduate students in those courses; second, to review the eight other syllabi for courses in the Archives and Public History program and suggest changes where needed; third, to arrange at least four digital history internships at cultural heritage institutions in the New York area to enable graduate students to gain significant experience working on digital history projects; and fourth, to create an entirely new *Advanced New Media* course.

The very existence of the Digital History Across the Curriculum project at NYU might be considered evidence in support of the claim that the field of digital humanities has ignored curricular reform, perhaps at its peril, perhaps thus enclaving itself. As the grant proposal points out, the NYU Archives and Public History program is one of only a few archival studies programs based in a history department rather than in a library and/or information science department, and this location in the humanities seems to have hampered the program's ability to adapt organically:

> Joseph M. Turrini, an archival educator at Auburn University, correctly observed in a May 2007 *AHA Perspectives* article that most history-based archival education programs have failed to adjust to changing professional expectations, especially 'the expansion of specialized archival courses and the increased technological expectations of archivists.' In truth, public historians, history departments, and humanities-based archival training programs have largely lagged behind their information science colleagues. George Mason University offers an M.A. in Applied History with a new media and technology emphasis and has emerged as a leader in the field, but few other institutions have followed along. Most programs offer isolated 'new media' courses at best. Indeed, recent curricular surveys of the public history field contain virtually no discussion of technology or digital issues. This appears particularly puzzling since these studies also document the fact that employers expect program graduates to possess precisely the blend of technological, collaborative, and administrative skills that immersion in digital history might provide. (Wosh, 2007)

For 'applied' humanities fields such as archival studies, documentary editing, and public history, the need for digital curricular reform or reinvention is obvious, if not always easy, placed as they are within humanities departments that rightly value critical inquiry for its own sake. It may be the case that digital humanities can only be transformative in humanities fields with just such an applied emphasis; even our project confines itself to revising the curriculum of a single track within a large history department, rather than the broader curriculum of the history department itself.

Yet one of the most promising and invigorating characteristics of digital humanities as a field is that it combines an emphasis on applied professional skills with the same deep respect for intellect, ethics, and emotion that animates the traditional humanities. Will there ever be an initiative to revise the entire curriculum of a History department, an English department, a French department, a Philosophy department to incorporate the issues and insights of the digital humanities? Does such an initiative seem unlikely, unnecessary, unimaginable? If so, then perhaps the digital humanities is quite properly segregated from humanities disciplines, and perhaps we can continue talking among ourselves.

References

Smith, M. N. (2008). Enclaves: Perils and Possibilities. Unpublished conference paper. *Digital Humanities and the Disciplines*. Rutgers University, New Brunswick, NJ, October 2008.

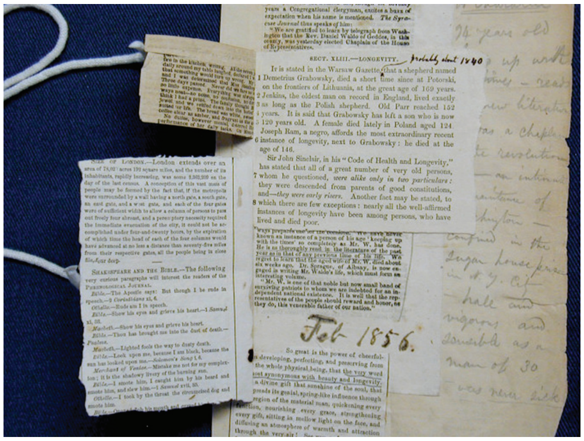
Turrini, J. M. (2007). The Historical Profession and Archival Education. *AHA Perspectives*, 45(5). http://www.historians.org/perspectives/issues/2007/0705/0705vie2.cfm (accessed 14 November 2008).

Wosh, P. J. (2007). Digital History across the Curriculum. Unpublished grant proposal. New York University: National Historical Publications and Records Commission.

Manuscript Annotations in Space and Time

Erica Fretwell
Duke University
enf3@duke.edu

Reading a book or manuscript involves four dimensions: width, length, depth and time, since flipping the page is a temporal act, with a before and an after. As scholars and publishers increasingly move towards digital remediation of literary archives, how do we digitally render 4D objects in a medium that is missing at least the third one, and text encoding standards that deprecate the role of the page in shaping meaning? Walt Whitman's annotations written on nineteenth-century books, magazines and newspapers, provide a special opportunity to begin to explore this question. A particularly unusual class of documents that Whitman created has posed difficulties for electronic rendering. At The Walt Whitman Archive, where I am a project manager, we call them "flipbooks," to denote a text that, while supported on a single sheet is, in essence, a scrapbook with multiple leaves glued on top of each other. That is, "flipbooks" are documents that are annotated and then layered on top of each other so that one can flip each clipping or page over:



years a Congregational clergyman, excites a buzz of expectation when his name is mentioned. *The Syracuse Journal* thus speaks of him:

"We are gratified to learn by telegraph from Washington that the Rev. Daniel Waldo of Geddes, in this county, was yesterday elected Chaplain of the House of Representatives.

SECT. XLIII.—LONGEVITY. Probably about 1840

It is stated in the Warsaw Gazette, that a shepherd named
1 Demetrius Grabowsky, died a short time since at Potorski, on the frontiers of Lithuania, at the great age of 169 years.
2 Jenkins, the oldest man on record in England, lived exactly
3 as long as the Polish shepherd. Old Parr reached 152
4 years. It is said that Grabowsky has left a son who is now
5 120 years old. A female died lately in Poland aged 124. Joseph Ram, a negro, affords the most extraordinary recent
6 instance of longevity, next to Grabowsky: he died at the age of 146.

Sir John Sinclair, in his "Code of Health and Longevity," has stated that all of a great number of very old persons,
7 whom he questioned, *were alike only in two particulars:* they were descended from parents of good constitutions, and—*they were early risers.* Another fact may be stated, to
8 which there are few exceptions: nearly all the well-affirmed instances of longevity have been among persons, who have lived and died poor.

ways prepares one for the occasion. We have never known an instance of a person of his age 'keeping up with the times' so completely as Mr. W. has done. He is as thoroughly read in the literature of the past year as in that of any previous time of his life. We regret to learn that the aged wife of Mr. W. died about six weeks ago. Dr. Sprague, of Albany, is now engaged in writing Mr. Waldo's life, which must form an interesting volume.

"Mr. W. is one of that noble but now small band of surviving patriots to whom we are indebted for an independent national existence. It is well that the representatives of the people should reward and honor, as they do, this venerable father of our nation."

Feb 1856.

So great is the power of cheerfulness in developing, perfecting, and preserving from the whole physical being, that the very word most synonymous with beauty and longevity. a divine gift that sunshine of the soul, that spreads its genial, spring-like influence through region of the material man, quickening every function, nourishing every grace, strengthening every gift, sitting in mellow light on the face, and diffusing an atmosphere of warmth and attraction through the very air!

SIZE OF LONDON.—London extends over an area of 78,02[illegible] acres 122 square miles, and the number of its inhabitants, rapidly increasing, was some 3,262,209 on the day of the last census. A conception of this vast mass of people may be formed by the fact that, if the metropolis were surrounded by a wall having a north gate, a south gate, an east gate, and a west gate, and each of the four gates were of sufficient width to allow a column of persons to pass out freely four abreast, and a peremptory necessity required the immediate evacuation of the city, it could not be accomplished under four-and-twenty hours, by the expiration of which time the head of each of the four columns would have advanced at no less a distance than seventy-five miles from their respective gates, all the people being in close file, four deep.

SHAKESPEARE AND THE BIBLE.—The following very curious paragraphs will interest the readers of the PHRENOLOGICAL JOURNAL.

Bible.—The Apostle says: But though I be rude in speech.—2 *Corinthians* xi, 6.

Othello.—Rude am I in speech.

Bible.—Show his eyes and grieve his heart.—1 *Samuel* xi, 33.

Macbeth.—Show his eyes and grieve his heart.

Bible.—Thou has brought me into the dust of death.—*Psalms.*

Macbeth.—Lighted fools the way to dusty death.

Bible.—Look upon me, because I am black, because the sun has looked upon me.—*Solomon's Song* i, 6.

Merchant of Venice.—Mistake me not for my complexion; it is the shadowy livery of the burning sun.

Bible.—I smote him, I caught him by his beard and smote him, and slew him.—1 *Samuel* xvii, 35.

Othello.—I took by the throat the circumcised dog and smote him.

94 years old
up with
times – read
new literature
was a Chaplain
the revolution
— an intimate
acquaintance of
Washington —
confined in the
Sugar house prison
in N. Y. City
hale and
vigorous and
sensible as a
man of 30
was never sick

In this presentation I will discuss how the questions of the third and fourth dimensions of these documents informed our design of interfaces for encoding, searching, and browsing digital surrogates of them.

With the help of an NEH grant in the Digital Humanities, the *Whitman Archive* created an interface that attempts to maximize audience and utility while addressing theoretical issues in the representation of layered documents.

We have created a set of software technologies and encoding practices that allow for the tagging, displaying, and searching of static documents that mix print, manuscript, and visual images—documents such as printed texts or images bearing handwritten annotations. These technologies include a suggested approach for encoding coordinates in XML transcriptions so that search engines can visually display results of user searches for manuscript words and phrases; web-based software for linking XML editing programs to an image display to allow encoders to relate bitmap images to XML text; and model stylesheets capable of displaying transcriptions of annotated documents together with digital images of those documents. We have kept Peter Robinson's warnings about the tendency of previous markup interfaces to be difficult to use—based on his experiences with his own software, Collate—in mind. For example, the ARCHway Project features a suite of powerful tools for relating texts to images and for capturing multiple hierarchies. Our software suite, following on the example of the ARCHway Project, is designed to be simple enough to be used by transcribers with little familiarity with information encoding and portable enough to work in multiple computing environments for widely different kinds of archival projects. Unlike ARCHway, however, it is web-based, and allows encoders to mark space as a structural entity.

Creating these interfaces has raised important questions. What exactly constitutes marginalia, and how would one render it digitally? To what extent are writing and reading, both in digital and non-digital media, spatial acts? Whitman's marginalia reveals his literary influences, how he is bound in time to writers that precede him; but the spatial element is significant as well, since his texts take root in the fertile marginal medium of theirs. His practice of pasting such documents together into unforeseen and reconfigurable combinations, with deliberately motile hierarchies, brings to mind Hayden White's argument that form not only reveals content, but can be content itself. It also reconfigures, as the work of Marta Werner on the "radical scatter[er]" Emily Dickinson suggests, what it means for scholars today to theorize peripheries. Hence, the theoretical orientations and practical implications of the archive's interface are not bound specifically to Whitman, but apply across literary studies and digital humanities. In creating a relationship between encoding practices and interactive design for search and browsing, we have followed a path suggested by, among others, Johanna Drucker, who asks if, "rather than think[ing] about simulating the way a book *looks*, we might consider extending the ways a book *works* as we shift into digital instruments."

Building an interface to allow for visualized search returns of manuscript words in images meant creating a coordinate system for manuscript tagging that allows for searches or deformations based on the location of a word in relative space, not just for specific terms or entities. But building one that makes possible an approach to how Whitman's flipbooks *work*, in Drucker's terms, also meant making alterations to the TEI P5 approach to XML-based markup, and in particular, to TEI's insistence that pages do not constitute intellectual structures. This presentation will briefly demonstrate our interface and discuss the logic of its design, focusing on the theoretical implications and practical potential of an approach that emphasizes layered spatial relations of text and images. I will discuss the logic of the "surface and zone" markup recommended for indicating such relations, and discuss why we found it unsuited to handling Whitman's annotated flipbooks. Briefly, I will outline our practical response, using a coordinate-tagging system, to such markup, which allows page space to function as an intellectual structure. To experiment with recreating the way temporal flexibility affects intellectual hierarchies in these documents, I will argue, not only introduces a new kind of activity, a new domain of interpretation, into editorial work, but also posits this kind of work as fundamentally multidimensional.

Bibliography

Bradley, Matthew. "GladCAT: An online catalogue of the books and reading of William Ewart Gladstone at St Deiniol's Library." St. Deiniols Library Web Site, January 2009. Accessed 1 Mar 2009. <http://www.st-deiniols.com/cms/goto.asp?id=142>

Dekhtyar, Alex, and Ionut Emil Iacob, "A Framework for Management of Concurrent XML Markup." Data and Knowledge Engineering 52.2 (2005): 185–215.

Drucker, Johanna. "The Virtual Codex from Page Space to E-space." Companion to Digital Humanities. Eds. Ray Siemens, John Unsworth, and Susan Schreibman. Oxford, UK: Blackwell, 2004. 198-217. <http://www.digitalhumanities.org/companion/>

Kiernan, Kevin, et al. "The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning." Literary and Linguistic Computing 20, Suppl 1 (2005): 69-88.

McGann, Jerome. "Marking Texts of Many Dimensions." Companion to Digital Humanities. Eds. Ray Siemens, John Unsworth, and Susan Schreibman. Oxford, UK: Blackwell, 2004. 198-217. <http://www.digitalhumanities.org/companion/>

Paul Dyck and Stuart Williams, “Toward an Electronic Edition of an Early Modern Assembled Book,” Computing in the Humanities Working Papers A.44, (July 2008).

Robinson, Peter. “Current issues in making digital editions of medieval texts—or, do electronic scholarly editions have a future.” Digital Medievalist 1.1 (Spring 2005). <http://www.digitalmedievalist.org/journal/1.1/robinson/>

Text Encoding Initiative. P5: Guidelines for Electronic Text Encoding and Interchange. 2007. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/>

Werner, Marta L., ed. Emily Dickinson’s Open Folios: Scenes of Reading, Surfaces of Writing. Ann Arbor: U of Michigan P, 1995.

_____, ed. Radical Scatters: Emily Dickinson’s Fragments and Related Texts, 1870-1886. Ann Arbor: U of Michigan P, 1999. <http://www.hti.umich.edu/d/dickinson/ >

Witt, Andreas, et al. “Unification of XML Documents with Concurrent Markup.” Literary and Linguistic Computing 2005 20.1: 103-116.

Supporting the Creation of Scholarly Bibliographies by Communities through Social Collaboration

Hamed Alhoori
Texas A&M University, USA
alhoori@tamu.edu

Omar Álvarez
Texas A&M University, USA
aomar@tamu.edu

Miguel Muñiz
Texas A&M University, USA
apresam@gmail.com

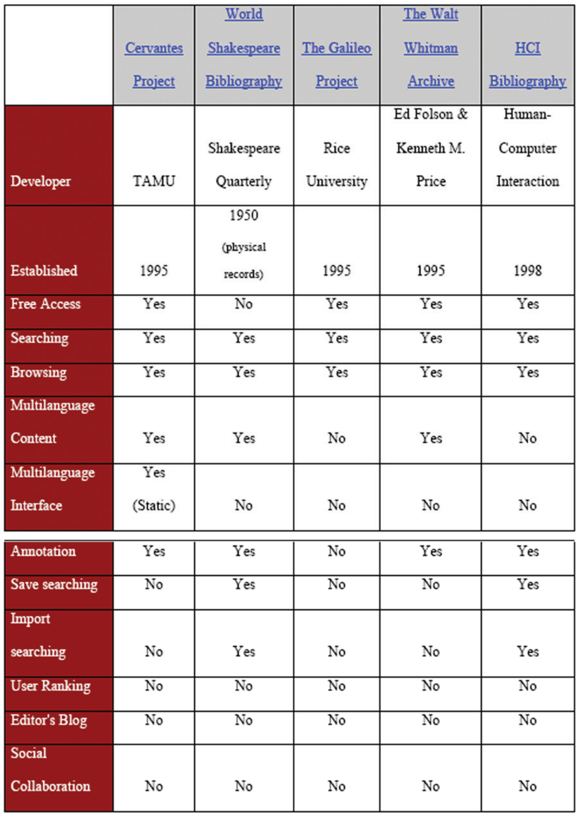
Richard Furuta
Texas A&M University, USA
furuta@cs.tamu.edu

Eduardo Urbina
Texas A&M University, USA
e-urbina@tamu.edu

Background

Many digital humanities projects maintain online bibliography digital libraries (BDLs) that support diverse users in locating a variety of references. The Cervantes Project (CP) bibliography aims to represent the best resources about Cervantes published since 1605 and is drawn from many multilingual sources. The current CP bibliography gathering and filtering process is carried out by sets of contributors: the expert editors, the reviewers, and the authorized international collaborators. Delays, possibly months, can result from the filtering process and also from the process of uploading the new publications into the BDL, which is separate from the gathering and filtering process. In addition, the ability to find new entries online is limited. Current bibliographic search engines show a limited scope of coverage on literature. There is no single resource that handles the entire 2.5 million articles that emerge yearly from the 25,000 peer-reviewed journals (Harnad, S. et al., 2008), so these engines access only a fraction of the literature (Hull, D. et al., 2008).

We compared various humanities BDL’s main supported features. Table 1 summarizes the main outcomes. Note that the majority of these BDLs do not take advantage of the social collaboration of Web 2.0.



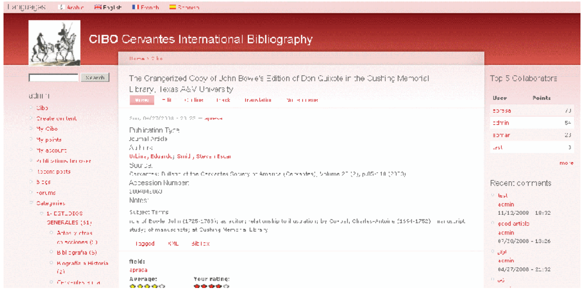
	Cervantes Project	World Shakespeare Bibliography	The Galileo Project	The Walt Whitman Archive	HCI Bibliography
Developer	TAMU	Shakespeare Quarterly	Rice University	Ed Folson & Kenneth M. Price	Human-Computer Interaction
Established	1995	1950 (physical records)	1995	1995	1998
Free Access	Yes	No	Yes	Yes	Yes
Searching	Yes	Yes	Yes	Yes	Yes
Browsing	Yes	Yes	Yes	Yes	Yes
Multilanguage Content	Yes	Yes	No	Yes	No
Multilanguage Interface	Yes (Static)	No	No	No	No
Annotation	Yes	Yes	No	Yes	Yes
Save searching	No	Yes	No	No	Yes
Import searching	No	Yes	No	No	Yes
User Ranking	No	No	No	No	No
Editor's Blog	No	No	No	No	No
Social Collaboration	No	No	No	No	No

Table 1. Various Humanities BDL supported features

This paper's premise is that social collaboration with the right level of moderation can support and reduce the costs of creating a scholarly bibliography by benefiting from the "wisdom of the crowds" (Surowiecki, J., 2004), while ensuring the accuracy of the bibliography. This could lead researchers to needed and interesting resources in better time. We have experimented with this issue by implementing a set of functionalities built on the drupal CMS. We have tested them on a group of CP users from different countries who use a variety of languages to gather, share, annotate, rank and discover academic literature (Fig. 1). We report on these initial experiments in the remainder of this paper.

Personalization

Zotero, Mendeley and Papers are personal reference management tools. However they do not include social collaboration features. We implemented a personal facility named *MyCibo* (Fig. 2), where users can save or edit their references, personal pages, and blogs with the ability to make them private or public. They can import and export in EndNote tagged, XML, RIS, and BibTeX formats and manually connect related publications.



The Grangerized Copy of John Bowle's Edition of Don Quixote in the Cushing Memorial Library, Texas A&M University

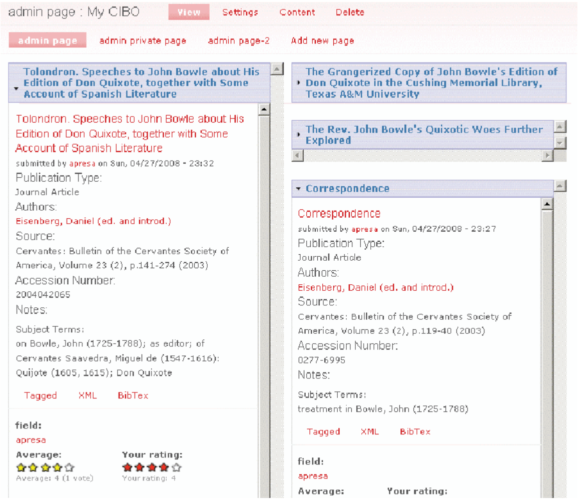
Sun, 04/27/2008 - 23:22 — apresa

Publication Type:
Journal Article
Authors:
Urbina, Eduardo; Smith, Steven Escar
Source:
Cervantes: Bulletin of the Cervantes Society of America (Cervantes), Volume 27 (2), p.85-110 (2007)
Accession Number:
2004042065
Notes:
Subject Terms:
role of Bowle, John (1725-1788); as editor; relationship to illustration by Coypel, Charles-Antoine (1694-1752); manuscript study; of manuscripts; at Cushing Memorial Library

Tagged XML BibTex

field:
apresa
Average:
Your rating:

Fig. 1 Screenshot from the main CIBO interface



Tolondron. Speeches to John Bowle about His Edition of Don Quixote, together with Some Account of Spanish Literature

Tolondron. Speeches to John Bowle about His Edition of Don Quixote, together with Some Account of Spanish Literature
submitted by apresa on Sun, 04/27/2008 - 23:32
Publication Type:
Journal Article
Authors:
Eisenberg, Daniel (ed. and introd.)
Source:
Cervantes: Bulletin of the Cervantes Society of America, Volume 23 (2), p.141-274 (2003)
Accession Number:
2004042065
Notes:
Subject Terms:
on Bowle, John (1725-1788); as editor; of Cervantes Saavedra, Miguel de (1547-1616): Quijote (1605, 1615); Don Quixote

Tagged XML BibTex

field:
apresa
Average:
★★★★☆
Average: 4 (1 vote)
Your rating:
★★★★☆
Your rating: 4

The Grangerized Copy of John Bowle's Edition of Don Quixote in the Cushing Memorial Library, Texas A&M University

The Rev. John Bowle's Quixotic Woes Further Explored

Correspondence

Correspondence
submitted by apresa on Sun, 04/27/2008 - 23:27
Publication Type:
Journal Article
Authors:
Eisenberg, Daniel (ed. and introd.)
Source:
Cervantes: Bulletin of the Cervantes Society of America, Volume 23 (2), p.119-40 (2003)
Accession Number:
0277-6995
Notes:
Subject Terms:
treatment in Bowle, John (1725-1788)

Tagged XML BibTex

field:
apresa
Average:
Your rating:

Fig. 2 Administrator MyCibo

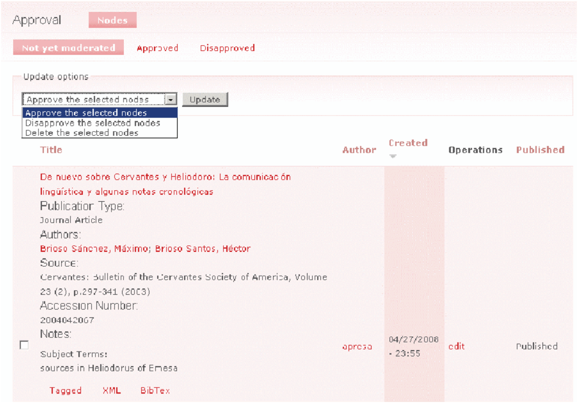
Social technologies applied to bibliographies

Social Bookmarking

Most online libraries and bibliographies provide one way learning, in that they provide services to the users, while prohibiting users from contributing. This results in a huge loss of knowledge and almost a freezing of storage rather than active libraries. The current state of the art is moving toward two way learning, where the users can both benefit from the available knowledge and contribute to it. (Hendry, D.G. et al. 2006b) mentioned an 'amateur bibliography' that is collected by nonprofessionals and falls short of the standards of a professional bibliography. Although large amount of references could be collected in a short span of time, resulting issues such as redundancy, spam, phantom author names, and phantom citations are not a good sign of scholarly research (Jacso, P . 2008).

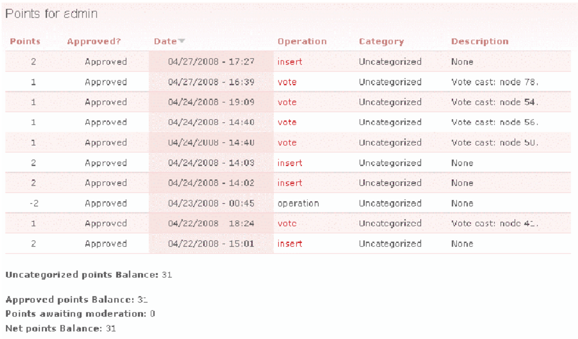
Unlike some general online reference management software such as CiteULike and Connotea that are based on the concept of non-moderated social bookmarking, we considered the previous issues and the need for an accurate bibliography. To get this done, light moderating and authenticating of the users contributions to the CP bibliography is provided, aiming to reach the scholarly moderated bibliography (Hendry, D.G. et al. 2006a).

Users were given ranks according to their scholarly or contribution level. For example, well known scholars got higher ranks so that they could contribute directly without moderating their contributions. New users' public contributions will be entered into a queue waiting for an approval from a moderator. Users who contribute with relevant and accurate contributions would mean that they are most likely experts in their area, and were given points, which promote their ranking and editing permissions. We believe this provides accuracy without losing the benefits of collaboration. Fig. 3 shows how to process the queued publications and Fig. 4 shows points gained by an administrator after several entries. Editors can revert to any previous revision in case there is need (Fig. 5).


Approval
Nodes
Not yet moderated
Approved
Disapproved
Update options
Approve the selected nodes
Update
Approve the selected nodes
Disapprove the selected nodes
Delete the selected nodes

	Title	Author	Created	Operations	Published
□	De nuevo sobre Cervantes y Heliodoro: La comunicac ón lingüística y algunas notas cronológicas Publication Type: Journal Article Authors: Brioso Sánchez, Máximo; Brioso Santos, Héctor Source: Cervantes: Bulletin of the Cervantes Society of America, Volume 23 (2), p.297-341 (2003) Accession Number: 2004042007 Notes: Subject Terms: sources in Heliodorus of Emesa Tagged XML BibTex	apresa	04/27/2008 - 23:55	edit	Published

Fig. 3 Process queued publications



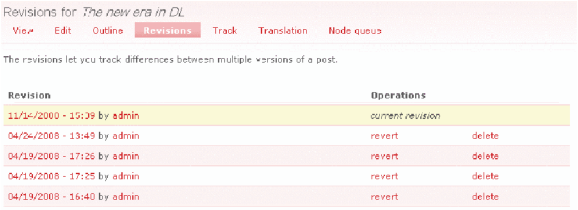
Points for admin

Points	Approved?	Date	Operation	Category	Description
2	Approved	04/27/2008 - 17:27	insert	Uncategorized	None
1	Approved	04/27/2008 - 16:39	vote	Uncategorized	Vote cast: node 78.
1	Approved	04/24/2008 - 19:09	vote	Uncategorized	Vote cast: node 54.
1	Approved	04/24/2008 - 14:40	vote	Uncategorized	Vote cast: node 56.
1	Approved	04/24/2008 - 14:40	vote	Uncategorized	Vote cast: node 50.
2	Approved	04/24/2008 - 14:03	insert	Uncategorized	None
2	Approved	04/24/2008 - 14:02	insert	Uncategorized	None
-2	Approved	04/23/2008 - 00:45	operation	Uncategorized	None
1	Approved	04/22/2008 18:24	vote	Uncategorized	Vote cast: node 41.
2	Approved	04/22/2008 - 15:01	insert	Uncategorized	None

Uncategorized points Balance: 31

Approved points Balance: 31
Points awaiting moderation: 0
Net points Balance: 31

Fig. 4 Detailed view of editors points



Revisions for *The new era in DL*

The revisions let you track differences between multiple versions of a post.

Revision	Operations	
11/14/2000 - 15:39 by admin	*current revision*	
04/24/2008 - 13:49 by admin	revert	delete
04/19/2008 - 17:26 by admin	revert	delete
04/19/2008 - 17:25 by admin	revert	delete
04/19/2008 - 16:40 by admin	revert	delete

Fig. 5 Revisions for a publication

We allowed the researchers to share and discover academic literature without worrying about inaccurate bibliographic data. They can discover what the warm topics are in the research field and what is significant to other researchers by viewing what other researchers read and tag. Hence, they can know the related researchers with similar interest that they can network with. Social collaboration is also important for papers that are not available electronically for various reasons and may loss their presence in the research community.

Social Tagging

Del.icio.us and Digg are of the best and fastest growing social bookmarking sites that use a folksonomy tagging. However, inaccurate and misleading tags are common in such open environments, which cannot be accepted in research communities. This is a classic Web 2.0 problem: it's hard to aggregate the wisdom of the crowd without aggregating their inexperience or madness as well (Torkington, N. 2006).

We prevent these effects by using social tagging with light moderation and users privileges upgrading. This provides us with a better quality tags than we would get if we just accepted all the beginners' tags; these users may want to contribute to the scholarly community initially but may loss their interest later on. We allowed the users to create their own tags and reuse the previously entered tags by them or other users using the AJAX technology, which allowed us to provide auto-complete tags in real time.

Social review and comments

There are different types of comments: approving, disapproving, or just summarizing the resource. We implemented a feedback environment to build an active online research community. It provides social reviews and comments from the users where the authors can interact with and answer their questions.

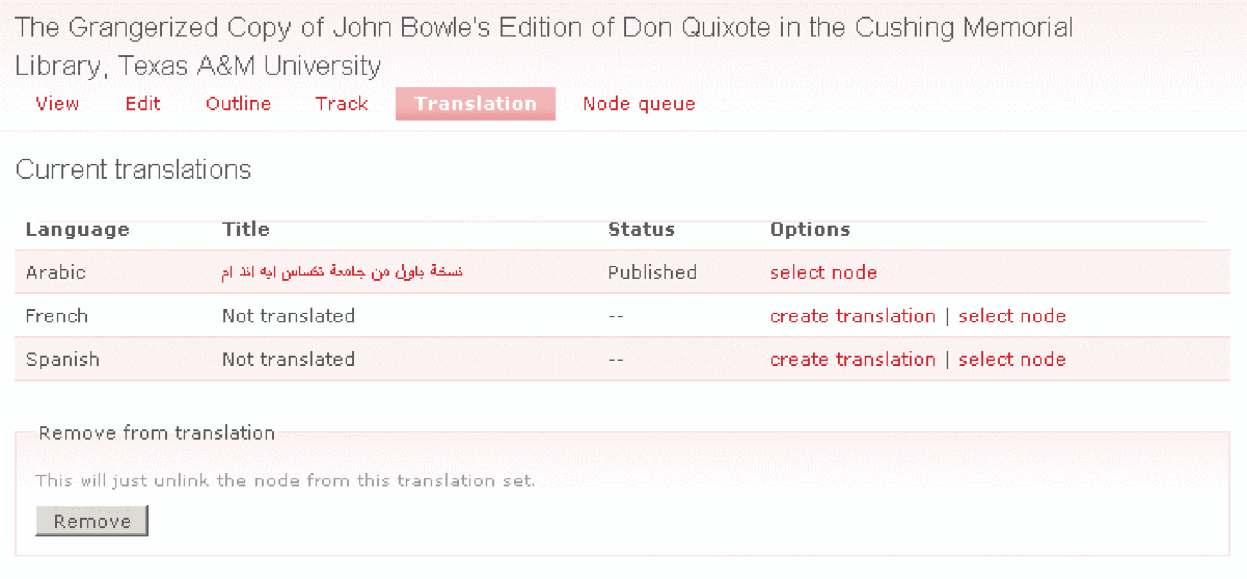
Multilanguage Capability

As digital libraries expand their audience and content scope, there is an increasing need for resources and access tools to those resources in a variety of languages (Larson, R.R. et al., 2002). Even for polyglot users, there is a preference to use maternal language interfaces in order to accelerate searching and browsing process, preferring the language of the interface to match the language of the content as well (Keegan, T . and Cunningham, S., 2008). Hence, the CP international scope requires the inclusion of content and system functionalities in multiple languages. Based on the statements presented, we provided a translation capability for interface elements (localization) and for content (internationalization). We analyzed different translation strategies such as using Web content (Wang, J. et al., 2004), documents in multiple languages (Nie, J.Y. et al., 1999), and some available APIs. After testing common searching phrases and sample texts in our content domain in three different languages (English, Spanish, and Arabic), we decided to use the Google AJAX Language API because of its detection and translation capabilities.

Users can choose the preferred available language at any moment while using the system. This choice will translate the interface to that language and would select only the content with that language. Bibliographic data can be entered in a language and then translated to a new language or linked to an existing bibliographic data or publications in other languages (Fig. 6). Users' comments and annotations can be translated to other languages, allowing users to comment and discuss in their preferred language (Fig. 7). The testing outcomes showed us acceptable translation results.

Ranking

Bibliography ranking has been used as a way to give users a confident Top-N resource from the searching results. A normal user only reads the first, second, or third page of results. Citations and references have been used as a way to rank bibliography resources (Larse, B. et al., 2002, Larse, B. et al., 2006, Yang, K. et al. 2007). Citation-based methods deal with complex issues such as biased or self-citations, hard to detect positive or negative citations, multiple citations formats difficult to handle by computer programs, unfair consideration of new papers, venues not considered. (Yan, Su, et al. 2007) propose a seed-based measure (considering top-venues and venues' authors relevance) and the browsing-based measure (considers user's behavior) to rank academic venues. However, the authors-seed needs to be updated frequently to reconsider new relevant authors. We used a hybrid approach. We allowed the users with higher ranking to rate the publications and retrieve the publications that got a vast amount of approved reviews and comments since that would mean that they are hot topics.


The Grangerized Copy of John Bowle's Edition of Don Quixote in the Cushing Memorial Library, Texas A&M University

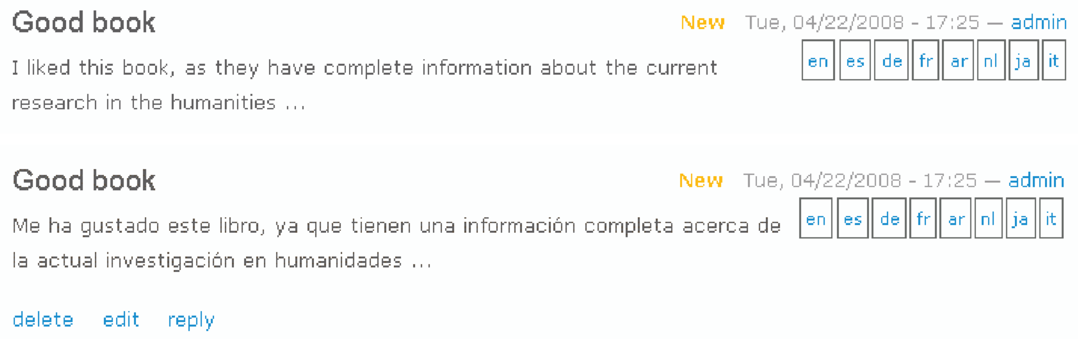
Current translations

Language	Title	Status	Options
Arabic	نسخة باول من جامعة تكساس ايه اند ام	Published	select node
French	Not translated	--	create translation | select node
Spanish	Not translated	--	create translation | select node

Remove from translation

This will just unlink the node from this translation set.

Fig. 6 Current publication translations


Good book

New Tue, 04/22/2008 - 17:25 — admin

I liked this book, as they have complete information about the current research in the humanities ...

Good book

New Tue, 04/22/2008 - 17:25 — admin

Me ha gustado este libro, ya que tienen una información completa acerca de la actual investigación en humanidades ...

Fig. 7 English comment translated to Spanish comment

Discussion and Future work

Our initial experimental results indicate that using an online social collaboration would improve the quality, quantity and usage of scholarly bibliography. Furthermore, it would help in building bridges among the international researchers and facilitate scholarly collaboration.

We intend to automate some portions of the moderating process and evaluate the reviews and comments (positive or negative) by identifying and interpreting annotations patterns and semantic to give some relevance weight to each source which would help also in the ranking.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0534314.

References

CiteULike. Available at:
http://www.citeulike.org (Accessed October 2008).

Connotea. Available at: www.connotea.org (Accessed October 2008).

Delicious. Available at: http://delicious.com/ (Accessed August 2008).

Digg. Available at: http://digg.com/ (Accessed August 2008).

Drupal. Available at: http://drupal.org/ (Accessed April 2008).

Google AJAX Language API, Available at: http://code.google.com/apis/ajaxlanguage/. (Accessed April 2008.)

Harnad, S. et al. (2008) The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update. Serials review, 34 (1). pp. 36-40.

Hendry, D.G. et al. (2006a). Hotlist or Bibliography? A Case of Genre on the Web, hicss,pp.51b, *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, p.51.2, January 04-07, 2006.

Hendry, D.G. et al. (2006b). Collaborative bibliography, *Information Processing and Management: an International Journal*, v.42 n.3, p.805-825, May 2006.

Hull, D. et al. (2008) Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web. PLoS Comput Biol 4(10): e1000204. doi:10.1371/journal.pcbi.1000204.

Jacso, P. (2008). Testing the Calculation of a Realistic *h*-index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster. Library Trends 56.4 (2008): 784-815. Project MUSE.

Keegan, T. and Cunningham, S. (2008). Language Preference in a Bi-language Digital Library, *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, Denver Colorado, USA, 2005.

Larson, R.R. et al. (2002). Harvesting Translingual Vocabulary Mappings for Multilingual Digital Libraries, *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries,* Portland Oregon, USA, 2002.

Larse, B. et al. (2002). The Boomerang Effect: Retrieving Scientific Documents via the Network of References and Citations, *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland, 2002.

Larse, B. et al. (2006). Using Citations for Ranking in Digital Libraries, *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, Chapel Hill, NC, USA, 2006.

Mendeley. Available at: http://www.mendeley.com/ (Accessed October 2008).

Nie, J.Y. (1999). Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web, *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, California, United States, Pages 74-81.1999. Papers. Available at: http://mekentosj.com/papers/ (Accessed October 2008).

Surowiecki, J. (2004). The Wisdom of the Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. 1st ed. New York: Doubleday.

Torkington, N. (2006). Digging the Madness of Crowds. http://radar.oreilly.com/archives/2006/01/digging-the-madness-of-crowds.html. (Accessed April 2008).

Wang, J. et al. (2004). Translating Unknown Cross-Lingual Queries in Digital Libraries Using a Web-based Approach, *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, Tucson, Arizona, USA; 2004.

Yan, S. et al. (2007). Toward Alternative Measures for Ranking Venues: A Case of Database Research Community*, Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, Vancouver, BC, Canada, 2007.

Yang, K. et al. (2007). CiteSearch: Next-generation Citation Analysis*, Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, Vancouver, British Columbia, Canada, 2007.

ZOTERO. Available at: http://www.zotero.org/ (Accessed October 2008).

LORE: A Compound Object Authoring and Publishing Tool for Literary Scholars

Anna Gerber
The University of Queensland
agerber@itee.uq.edu.au

Jane Hunter
The University of Queensland
jane@itee.uq.edu.au

This paper presents LORE (Literature Object Re-use and Exchange), a light-weight tool designed to enable scholars and teachers of literature to author, edit and publish OAI-ORE-compliant compound information objects that encapsulate related digital resources and bibliographic records. LORE provides a graphical user interface for creating, labelling and visualizing typed relationships between individual objects using terms from a bibliographic ontology based on the IFLA FRBR. After creating a compound object, users can attach metadata and publish it to a repository as an RDF graph, where it can be searched, retrieved, edited and re-used by others. LORE has been developed in the context of the Australian Literature Resource project (AustLit) and hence focuses on compound objects for teaching and research within the Australian literature studies community.

1. Introduction and Objectives

Within the discipline of literature research and teaching, the ability to relate disparate digital resources in a standardized, machine-readable format has the potential to add significant value to distributed collections of literary resources. Such compound objects can be used to track the lineage of derivative works which are based on a common concept, to relate objects around a common theme, or to encapsulate related digital resources for teaching purposes. For example, one might want to relate the original edition of *Follow the Rabbit-Proof Fence* to the illustrated edition, a radio recording and a digital version of the film – and to retrieve and present these resources, with their relationships visualized, regardless of their location. Our objective is to provide a software tool to enable such encapsulation and subsequent re-use and visualization, by building on the efforts of two previous digital library initiatives:

- The IFLA Functional Requirements for Bibliographic Records (IFLA, 1998)

- The OAI-Object Reuse and Exchange (OAI, 2008)

FRBR is a recommendation of the International Federation of Library Associations and Institutions (IFLA) to restructure catalogue databases to reflect the conceptual structure of information resources. It uses an entity-relationship model of metadata for bibliographic resources that supports four levels of representation: work, expression, manifestation and item. It also supports three groups of entities: products of intellectual or artistic endeavour (publications); entities responsible for intellectual or artistic content (a person or organisation); and entities that serve as subjects of intellectual or artistic endeavour (concept, object, event, and place).

The Open Archives Initiative Object Reuse and Exchange (OAI-ORE) is an international collaborative initiative, focusing on a framework for the exchange of information about Digital Objects between cooperating repositories, registries and services. OAI-ORE aims to support the creation, management and dissemination of the new forms of composite digital resources being produced by eResearch and to make the information within these objects discoverable, machine-readable, interoperable and reusable. Named Graphs (Jeremy, 2005) are endorsed by the OAI-ORE initiative as a means of publishing compound digital objects that clearly states their logical boundaries (Lagoze et al, 2007). They do this in a way that is discipline-independent, but that also provides hooks to include rich semantics, metadata, ontologies and rules. *Our hypothesis is that OAI-ORE Named Graphs provide the ideal mechanism for representing literary compound objects that encapsulate the entities and relationships expressed by the IFLA FRBR.*

To test this hypothesis, we are working with the Australian literature studies community through AustLit. AustLit is a non-profit collaboration between the National Library of Australia and twelve Universities. It provides the peak resource of bibliographic data for scholars undertaking research into Australian literary heritage and print culture history. The AustLit data model is also based on the IFLA FRBR (Kilner, 2005), making it ideal for evaluating LORE. Hence our core aims are to provide easy-to-use tools that can be seamlessly integrated within existing research practices through the AustLit Web Portal and that enable:

- the publishing of compound objects in open access repositories so they can be readily shared and re-used;

- the easy discovery and re-use of these compound objects through the attachment of simple metadata;

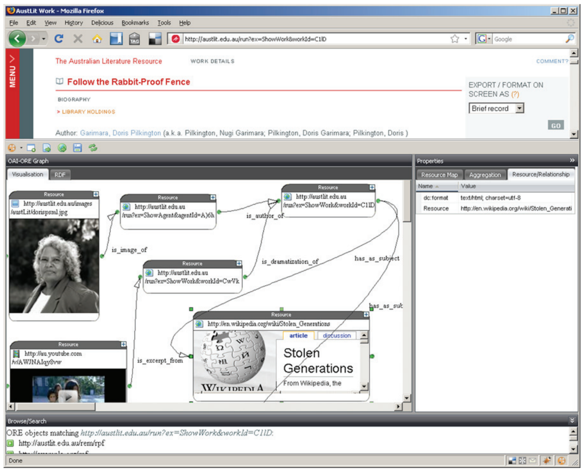
- the visualization of complex relationships between literary resources (including the lineage of derived intellectual products) through intuitive graphical user interfaces.

2. Related Work

A number of previous efforts have applied OAI-ORE to specific scientific disciplines to encapsulate experimental data and results. These include: FORSITE (2008), eChemistry (Van Noorden, 2008), UIUC (Cole, 2008) and SCOPE (Cheung et al, 2007). Although CULTOS (2003) uses RDF to represent multimedia and hypertext presentations for e-Humanities applications, it does not combine OAI-ORE and IFLA-FRBR to capture or label the precise relationships between entities. Also relevant is an overview of previous implementations and applications of IFLA FRBR, provided by Babeu (2008). A significant past focus of e-Humanities tools development has been on scholarly mark-up and annotation tools to attach interpretations to individual objects or parts of objects (e.g., paragraphs within an article). LORE takes the annotation paradigm a step further, enabling authors to annotate links between multiple resources with tags from an ontology.

3. Implementation and User Interface

LORE is implemented as a Mozilla Firefox extension using AJAX. The LORE tool stores and queries Named Graphs representing compound objects via web services on a Sesame 2 or Fedora repository. The types for intra-aggregation relationships as well as metadata terms for aggregated objects are specified via an OWL ontology, which is configured at start-up. Through examining all of the topic types and relationships from the AustLit database, we developed an OWL ontology which is based on IFLA FRBR, but extended to support additional relationships (e.g., between people).


AustLit Work - Mozilla Firefox
File Edit View History Delicious Bookmarks Tools Help
http://austlit.edu.au/run?ex=ShowWork&workId=C1lD
Google
MENU
The Australian Literature Resource
WORK DETAILS
COMMENT?
Follow the Rabbit-Proof Fence
BIOGRAPHY
> LIBRARY HOLDINGS
EXPORT / FORMAT ON SCREEN AS (?)
Brief record
GO
OAI-ORE Graph
Visualisation
RDF
Resource
http://austlit.edu.au/images/austLit/dorispsml.jpg
Resource
http://austlit.edu.au/run?ex=ShowAgent&agentId=A)6h
Resource
http://austlit.edu.au/run?ex=ShowWork&workId=C1lD
is_author_of
is_image_of
Resource
http://austlit.edu.au/run?ex=ShowWork&workId=CwVk
is_dramatization_of
has_as_subject
has_as_sub
Resource
http://au.youtube.com/v/AWJNAIqy8vw
is_excerpt_from
Resource
http://en.wikipedia.org/wiki/Stolen_Generations
article
discussion
Stolen Generations
From Wikipedia, the
Properties
Resource Map
Aggregation
Resource/Relationship
Browse/Search
Done

Author: Garimara, Doris Pilkington (a.k.a. Pilkington, Nugi Garimara; Pilkington, Doris Garimara; Pilkington, Doris)

Name	Value
dc:format	text/html; charset=utf-8
Resource	http://en.wikipedia.org/wiki/Stolen_Generati

ORE objects matching *http://austlit.edu.au/run?ex=ShowWork&workId=C1lD:*
http://austlit.edu.au/rem/rpf

Fig. 1. Compound object editing interface

LORE's editing interface displays OAI-ORE resource maps in a graphical form, as shown in Figure 1, as well as RDF/XML. In the graphical view, nodes represent the resources aggregated within the resource map and arcs represent typed relationships between them. Each graphical node contains an interactive preview of the resource that it represents, which can be collapsed to conserve screen space or resized to display more content. This allows users to view and interact with aggregated resources directly from within LORE rather than having to load them individually in the browser. Clicking on a node's identifier loads the resource in the top browser window.

Metadata about the OAI-ORE resource map, aggregation, and aggregated resources is displayed and can be added to or edited via the *Properties* panel on the bottom right-hand-side. The metadata terms that may be specified are those from OAI-ORE, Dublin Core (DC, 2008), DCMI Metadata Terms (DCMI, 2008), selected terms from FOAF (FOAF, 2007), and, datatype properties from the domain ontology. Relationship types are indicated by labels on the arcs, and can be changed by editing the properties or by selecting from the arc context menu, which is populated by the object properties from the domain ontology.

New resources to be added to the resource map are discovered via the main browser window. Clicking on the OAI-ORE logo in the status bar toggles the editor's visibility, so that the full window can be used for resource discovery, whilst the resource map being constructed remains accessible throughout the browsing session. A resource loaded in the browser can be added to the resource map via context menus or LORE's toolbar. The toolbar provides options for saving and loading compound objects stored in the RDF repository specified in the user preferences. Resource maps can also be discovered and loaded via the *Browse/Search* panel.

4. Discussion and Conclusions

The AustLit researchers with whom we have been collaborating have been overwhelmingly enthusiastic about this work. They particularly liked the interactive node previews, the direct integration of the editor with the browser and the ease with which they could customize the relationship types and metadata supported by the editor. They would like to see additional arc visualizations such as line decorators, arrows and colours or line styles to distinguish relationships, as well as support for bi-directional relationships.

Objects can be added to a compound object in the LORE editor if they can be loaded in the web browser. However this approach does not handle URIs identifying non-information resources well, and issues arise with non-persistent URLs and with identifying objects that exist within institutional repositories using local identifiers.

Because the IFLA FRBR is complex, it may be difficult for a literary scholar to apply appropriate metadata terms and relationship types from the ontology to relate resources. Strategies for addressing this issue could include adding more semantic checks to the UI to assist users in applying the ontology terms, or tailoring the domain ontologies based on community needs and understanding.

The on-going development and evaluation of LORE in the context of AustLit will provide an essential component of the cyber-infrastructure requirements of the Australian literary studies community, as well as literary scholars globally.

Acknowledgements

Aus-e-Lit is funded by DEST through the National eResearch Architecture Taskforce (NeAT), part of the National Collaborative Research Infrastructure Strategy (NCRIS).

References

AustLit. (2008). AustLit: The Australian Literature Resource. http://austlit.edu.au (accessed 11 November 2008).

Babeu, A. (2008). Building a "FRBR-Inspired" Catalog: The Perseus Digital Library Experience. http://www.perseus.tufts.edu/~ababeu/PerseusFRBRExperiment.pdf (accessed 11 November 2008).

Cheung, K., Hunter, J., Lashtabeg, A., Drennan, J. (2007). SCOPE - A Scientific Compound Object Publishing and Editing System, *3rd International Digital Curation Conference*, Washington DC.

Cole, T.W. (2008). OAI-ORE experiments at the University of Illinois Library at Urbana-Champaign. http://www.openarchives.org/ore/meetings/Soton/Cole-OAI-

ORE-Roll-Out-OR08.pdf (accessed 11 November 2008).

CULTOS. (2003). http://www.cultos.org/ (accessed 11 November 2008).

DC. (2008). Dublin Core Metadata Element Set, Version 1.1. http://dublincore.org/documents/dces/ (accessed 11 November 2008).

DCMI. (2008). DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/ (accessed 11 November 2008).

FOAF. (2007). FOAF Vocabulary Specification 0.91. http://xmlns.com/foaf/spec/ (accessed 11 November 2008).

FORSITE. (2008). http://foresite.cheshire3.org/ (accessed 11 November 2008).

IFLA. (1998). Functional requirements for bibliographic records (FRBR): Final report. http://www.ifla.org/VII/s13/frbr/frbr.pdf (accessed 11 November 2008).

Jeremy, J.C., et al. (2005). Named graphs, provenance and trust, *14th international conference on World Wide Web*. ACM Press, Chiba, Japan

Kilner, K. (2005). The AustLit Gateway and Scholarly Bibliography: A Specialist Implementation of the FRBR. *Cataloguing and Classification Quarterly*. 39:3/4.

Lagoze, C., Van de Sompel, H. (2007). Compound Information Objects: The OAIORE Perspective. http://www.openarchives.org/ore/documents/CompoundObjects-200705.html (accessed 11 November 2008).

OAI. (2008). Open Archives Initiative - Object Reuse and Exchange. http://www.openarchives.org/ore/ (accessed 11 November 2008).

Van Noorden, R. (2008). Microsoft Ventures into Open Access Chemistry. *Chemistry World*. http://www.rsc.org/chemistryworld/News/2008/January/29010803.asp (accessed 11 November 2008).

Gobineau and Tocqueville: The Curious Case of the Medical Metaphor in Corpus Stylistics

Joel Goldfield
Fairfield University
jgoldfield@mail.fairfield.edu

Corpus stylistics and statistical analysis of keywords can have a voice in a growing debate about the literary effect of metaphor between two nineteenth-century authors who also served as French government officials: Arthur de Gobineau (1816-1882) and Alexis de Tocqueville (1805-1859). In his prize-winning biography on Tocqueville, Hugh Brogan recently highlighted an aggressive response made 150 years before by the esteemed literary statesman to his former chief of staff. Tocqueville, Brogan argues, disagreed "...entirely with Gobineau's immoral ideas about human decadence" (592) and miscegeny, particularly as expressed in the latter's *L'Essai sur l'inégalité des races humaines*, whose first two volumes Tocqueville had read. In his defense, Gobineau, at that point a diplomat in France's Foreign Service, replied that, "he was no more immoral than a doctor who tells his patient that his disease is mortal" (592). Brogan then cites, in his own translation, a key paragraph from the published correspondence of July 30, 1856, from Tocqueville to Gobineau:

> I reply that if the act is not immoral in itself, it can only produce immoral or pernicious consequences. If my doctor came to me one morning to say, 'My dear sir, I have the honour to announce that you have a mortal illness, and as it affects your very constitution, I have the advantage of being able to add that there is absolutely no chance of saving you in any way,' I would first be tempted to knock the fellow down. (592)

To Tocqueville's follow-up thought that he would then hide under the blanket or perhaps, like Boccaccio's characters during the plague in Florence, indulge his whims or perhaps prepare himself for the afterlife before the inevitable occurred, Brogan adds a provocative footnote: "It might be interesting to know exactly when AT [Alexis de Tocqueville] read Gobineau's latest volumes. In the Foreword to the *Ancien Régime* he compares himself to a physician (OC II i 73)" (592).

With so much of Tocqueville's and Gobineau's literary and historical work online in the ARTFL database, corpus stylistics and statistical analysis may have something to offer our research on the vocabulary of Tocqueville's striking reply of July 1856, compared to the wording in his foreword to *L'Ancien régime et la révolution*. The first two volumes of Gobineau's *Essai* had been already been published in the summer of 1853. The first two occurrences of *médecin(s)*, both in the first two volumes released, and the fourth, in the last two volumes (1855) to which Brogan refers as "Gobineau's latest volumes," are simply part of a list of professions. The third, however, occurs in a discussion of Plato. Gobineau explains that either because of status attributable to his birth or because of circumstances, Plato finds himself in charge. Unfortunately, horrified by Athen's problems and hesitant to undertake any mission that might worsen them, Plato remains powerless to act. "Such people," Gobineau opines, "are doctors, not surgeons and, like Epaminondas and Philopoemen, they cover themselves in glory without fixing anything" (my translation). And so it would seem that doctors may be good only for diagnosing problems, not repairing the damage. Indeed, Gobineau had written to Tocqueville on March 20, 1856, "I am no more an assassin than the doctor who says that the end is near. I'm wrong or I'm right." All written in relatively close chronological succession, Gobineau's thoughts seem to imply that he is like Plato and like a doctor, intelligent enough to see the errors of mankind's ways, able to speak up to diagnose the ills, but powerless to solve any of them. How inconvenient! Explaining away any curative or corrective abilities that doctors may have is a false premise, perhaps appealing to a minority of popular complaints, and seems to serve as justification for the supposedly perceptive Gobineau to throw his hands in the air. Tocqueville, however, remains hopeful, despite the ill health that has been plaguing him for years: "I add that physicians, like the philosophes, are often wrong in their predictions, and I have seen more than one man condemned by them carrying himself quite well later while resenting the doctor who had needlessly frightened and discouraged him."

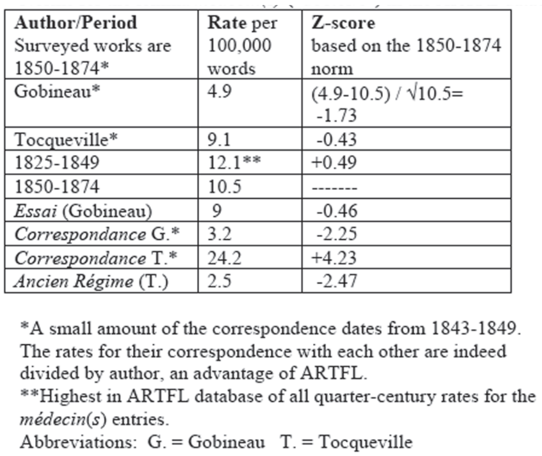
Research on the delivery patterns of letters between Gobineau and Tocqueville indicates the probability that enough time existed for Gobineau's letter of March 20, 1856, sent by diplomatic post from Teheran, to reach Tocqueville in Paris before the final weeks in May and the first two weeks in early June when he wrote or revised, respectively, the foreword to L'Ancien régime (see Gannett, p. 144). Tocqueville's work was not published until June 16, 1856 (see http://classiques.uqac.ca/classiques/De_tocqueville_alexis/ancien_regime/Ancien_regime.pdf).

"I therefore admit that in studying our former society in each of its parts, I have never entirely lost the new one from sight. I have not only wanted to see to what illness the patient had succumbed, but how he could have avoided dying. I have done as those doctors who, for each exhausted organ, try to surprise the laws of life. My goal has been to paint a picture that would be exact and that at the same time could be instructive" (my translation).

(*"J'avoue donc qu'en étudiant notre ancienne société dans chacune de ses parties, je n'ai jamais perdu entièrement de vue la nouvelle. Je n'ai pas seulement voulu voir à quel mal le malade avait succombé, mais comment il aurait pu ne pas mourir. J'ai fait comme ces médecins qui, dans chaque organe éteint, essayent de surprendre les lois de la vie. Mon but a été de faire un tableau qui fût strictement exact, et qui, en même temps, pût être instructif."*)

One might wonder whether Tocqueville's delicate health and Gobineau's repugnant preference for partitioned racial types as a way to stabilize mankind's existence might have led to added references to doctors and related terms. Judging by the data in the ARTFL FRANTEXT database, the entire exchange described above is, on the contrary, quite uncommon for these two authors as individuals. Tocqueville's use of the combination or cluster médecin(s)- vie(s)–organe(s) in the foreword of L'Ancien régime is one of only seven such clusters within the same sentence over the entire FRANTEXT database, from the early 1600's through 1992 in the 2,540 documents searched. And two of these seven were purely medically oriented.

The z-scores in Figure 1 suggest at least two somewhat surprising interpretations. First, Tocqueville, rather than Gobineau, seems to have the more medically oriented attitude toward politics and civilization. The z-score of +4.23 in Tocqueville's correspondence is statistically significant in its positive rate whereas Gobineau's rate of -2.25 is an inverse correlate, significantly negative. Second, the use of the lemma médecin in Tocqueville's Ancien régime, during whose writing he became increasingly ill with tuberculosis, is also negatively significant. This fact heightens the value of the cluster doctor(s)-life/lives-organ(s) (médecin(s)- vie(s)–organe(s)) described above. Returning to the original observation by Hugh Brogan, one might suppose that he and perhaps others who are scholars or otherwise frequent readers of Tocqueville's work are likely to be even more impressed by the physician/doctor metaphor. While the use of médecin(s) in L'Ancien régime does not attract attention for repetition compared to the norm of the quarter century, it certainly does for cognitive and statistical reasons in Tocqueville's and Gobineau's oeuvres . The bas-relief of the word's rarity in this work moves the word and its word cluster to the forefront, particularly in a foreword. Coupled with time and opportunity, it indeed seems likely that Tocqueville had a rebuttal to Gobineau on his mind as he penned the avant-propos shortly before his own father's death. It is no coincidence that in the next sentence after the medical metaphor, he writes of fathers, male virtues ("vertus mâles"), and faith in a cause as well as in ourselves, "la foi en nous-mêmes et dans une cause."



Author/Period Surveyed works are 1850-1874*	**Rate** per 100,000 words	**Z-score** based on the 1850-1874 norm
Gobineau*	4.9	(4.9-10.5) / √10.5= -1.73
Tocqueville*	9.1	-0.43
1825-1849	12.1**	+0.49
1850-1874	10.5	-------
Essai (Gobineau)	9	-0.46
Correspondance G.*	3.2	-2.25
Correspondance T.*	24.2	+4.23
Ancien Régime (T.)	2.5	-2.47

*A small amount of the correspondence dates from 1843-1849. The rates for their correspondence with each other are indeed divided by author, an advantage of ARTFL.
**Highest in ARTFL database of all quarter-century rates for the *médecin*(*s*) entries.
Abbreviations: G. = Gobineau T. = Tocqueville

Figure 1. Norms for the lemma médicin(s) ("doctor/s") in the ARTFL database

References

ARTFL. American and French Research on the Treasury of the French Language. U. of Chicago.

Brogan, Hugh (2006) *Alexis de Tocqueville: A Life*. New Haven: Yale.

Gannett, Robert T., Jr. (2003) *Tocqueville Unveiled*. Chicago & London: University of Chicago.

Gobineau, Arthur de (1853) *Essai sur l'inégalité des races humaines*, vols. 1 & 2. Paris: Pierre Belfond. ARTFL database.

---- (1855) *Essai sur l'inégalité des races humaines*, vols. 3 & 4. Paris: Pierre Belfond. ARTFL database.

Tocqueville, Alexis de (1959) *Oeuvres complètes. Tome II. L'Ancien régime et la révolution*. Originally published in 1856. Also see: see http://classiques.uqac.ca/classiques/De_tocqueville_alexis/ancien_regime/Ancien_regime.pdf.

--- *Oeuvres complètes. Tome IX. Correspondance d'Alexis de Tocqueville et d'Arthur de Gobineau*. Introduction by J.-J. Chevallier. Edited and annotated by Maurice Degros. Paris: Gallimard. Correspondence dates from 1843-1859.

Define Me: A Cognitive and Computational Approach to Critical Digital Identity Representation in Social Networking Applications

D. Fox Harrell
Georgia Institute of Technology
fox.harrell@lcc.gatech.edu

Daniel Upton
Georgia Institute of Technology
dupton3@gatech.edu

Ben Medler
Georgia Institute of Technology
benmedler@gatech.edu

Jichen Zhu
Georgia Institute of Technology
jichen.zhu@lcc.gatech.edu

Invoking a theoretical framework situated at the intersection of humanistic accounts of social identity construction, cognition linguistics research, and digital media technologies, the Advanced Identity Representation (AIR) Project develops theory and technology for users to represent complex, dynamic social identities in digital media such as virtual worlds and social networking sites. Here, we primarily present *DefineMe – Chimera*, a social networking application that uses a dynamic system of categorization and allows users to define each other through metaphorical projection. *DefineMe* is grounded in an interdisciplinary approach that articulates the shared socio-cognitive substrates beneath user representations ranging from user created profiles on social networking sites to avatars in virtual worlds. Secondarily, we present *Identity Share*, a social networking project developed using the *DefineMe* database structure that allows users to define identity categories, share profiles, and anonymously follow each other's web searching paths. The result of the projects is an early articulation of a spectrum of new user identity representations with foci upon group membership, utilization/creation of boundary infrastructures (Bowker & Star 1999; Lave & Wenger 1991), along with cognitive models of metonomy, metaphor, and visual imagery. (Hutchins 1996; Lakoff 1987)

1. Introduction

The Advanced Identity Representation (AIR) Project is the name given to the research endeavor in Fox Harrell's Imagination, Computation, and Expression (ICE) Lab/Studio investigating technology and theory to enable digital experiences that engage a richer range of social identity experiences than those found now in social networking, gaming, and virtual worlds software. We present *DefineMe – Chimera* and *Identity Share* as early steps toward this end. *DefineMe – Chimera* is a social networking application with a novel database system and a front-end Facebook web application. Users can label each other using self-defined predicates expressing their metaphorical similarities to various animals. These descriptions are used as a basis to construct and reconfigure categories on-the-fly as the database grows and to present chimera-like avatar characters to represent the user as composites of various iconic animal graphics. Though this project develops a whimsical, metaphorical model of user representation, the theoretical and technical underpinnings address issues such as coconstruction of identity categories between individuals, marginalization and centrality within identity categories, and the imaginative nature of identity in race, ethnicity, and gender critical contexts.

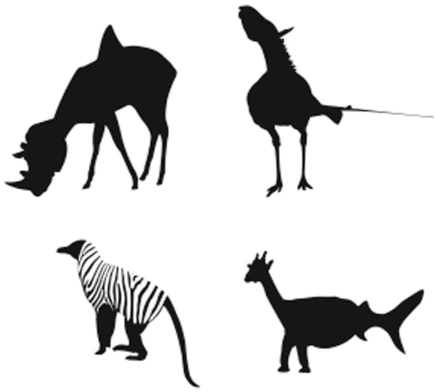


Fig. 1 Metaphorical animal blend avatars potentially generated by DefineMe – Chimera

As a second step toward enabling a new genre of digital media identity experiences, we present *Identity Share*, a critical web-based application that offers a balance between allowing users to author profiles with both self-defined and normative social categories, at the same time as allowing users to specify the relative importance of particular categories. *Identity Share* allows users to anonymously give others permission to follow their web searches, view their wishlists with various websites, and leave comments on their experiences of "(web)walking in another's shoes." The goal is not to connect users as friends, but rather to allow users to have the uncanny experience of viewing and sharing aspects of each others' needs, values, and desires. Together, *DefineMe: Chimera* and *Identity Share* exemplify early prototypes of the direction that AIR Project systems may take in tackling social identity phenomena in the future.

2. Theoretical Framework

The AIR Project draws on a hybrid approach to issues of categorization and classification. Central influential theories important to the AIR Framework are described below.

2.1 Cognitive Categorization, Metaphor, and Blending

The AIR approach has some of its roots in grounded in cognitive science theory (Lakoff 1987) which asserts that categorization is a matter of both human experience and imagination. George Lakoff's work in this area over two decades ago is well known and influential, yet to our knowledge it is a thread that has been underdeveloped with respect to issues of social identity construction in the critical modes robustly developed in cultural studies with the humanities, especially this approach has not been significantly applied to cases of digital identity representation (an exciting exception being the work of Otto Santa Ana on metaphorical bias in *Brown Tide Rising* (Santa Ana 2002)). Cognitive science research reveals categories as being (1) based on "the same neural and cognitive mechanisms that allow us to perceive and move around" (Lakoff & Johnson 1999), (2) distributed across members of a social group, external artifacts, and even time (Hutchins 1996, 2000), and (3) always situated in particular social and cultural contexts (Lave & Wenger 1991). Important for the purposes here, Lakoff describes a conceptual metaphor-based theory of how imaginative extensions of "prototype effects" result in several phenomena of social identity categorization that are useful for the AIR Project (Lakoff 1987). These phenomena include *representatives* (prototypes) or "best example" members of categories, *stereotypes* that indicate normal, but often misleading, category expectations, and more. Conceptual blending theory builds upon Gilles Fauconnier's mental spaces theory (Fauconnier 1985), elaborates insights from metaphor theory (Fauconnier 2006), and attempts to account for a wider range of semantic phenomena.

2.2 Sociology of Classification Infrastructures

The AIR Project is influenced by accounts of classification from sociology and science studies. In *Sorting Things Out* (Bowker & Star 1999), Geoffrey Bowker and Susan Leigh Star make the case that classification systems are necessary for information exchange and communication. The social challenges regarding classification systems arise from cases where tension exists between contexts, for examples, when one's self-conception differs from prevalent social stereotypes. Important tools for bridging between communities are "boundary objects," defined by Bowker and Star as objects that "inhabit several communities and also satisfy the informational requirements of each." The AIR Project develops what Bowker and Star term "boundary infrastructures." These are defined as "stable regimes managing multiple boundary objects, allowing the necessary information to be accessed by multiple communities." Also crucial from Bowker and Star, is the concept of "torque," the condition where biographies are "twisted in classification systems" to arrive at painful lived experiences. One poignant example Bowker and Star present is the schism between societal and self-perception and the disruptive movement between or misapplication of categories, especially for people labeled as "black" or "colored," in apartheid South Africa. The gap between self (or local community-based) definition of an individual's place in a classification system and hegemonically imposed definition of classifications, and disarming the negative results often arising from such phenomena, is a central to the critique performed by the systems highlighted in this paper. As opposed to computational identity applications that are based on standard, static classification systems, the dynamically configurable, imaginatively grounded AIR Project identity systems are boundary infrastructures that allow users to customize their user profiles and preferences for particular communities.

3. The AIR Framework

Based upon the cognitive and infrastructural approach above, and previous work in imaginative computational discourse and identity construction (Harrell 2007, 2008a), a brief summary of key aspects of the AIR Project's new constructs for implementing and analyzing computational identity follows.

3.1 Shared Technical Underpinnings of Identity Applications

A technical infrastructure-oriented means to compare computational identities is the first pillar the approach developed in this project. Various computational identity applications such as social networking sites, avatar creation systems for virtual worlds, and games are implemented using a limited and often overlapping set of techniques. Fig. 2 describes, at a high level, the components that comprise the majority of widely used computational identity technologies (Harrell 2008b).

The six components in Fig. 2 that commonly form the basis for avatar/character/profile construction can enable dynamic and contingent models of social identity in digital environments as described in (Gee 2003). Understanding the reciprocities and overlaps between the technical means by which users stage their identities across digital media forms can enable more powerful customizability and cross-community communication facility in social identity systems.

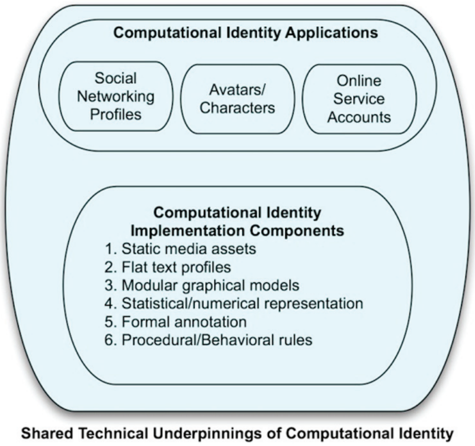

Computational Identity Applications
Social Networking Profiles
Avatars/ Characters
Online Service Accounts
Computational Identity Implementation Components
1. Static media assets
2. Flat text profiles
3. Modular graphical models
4. Statistical/numerical representation
5. Formal annotation
6. Procedural/Behavioral rules
Shared Technical Underpinnings of Computational Identity

Fig. 2 Shared Technical Underpinnings of Computational Identity Applications

3.2 Cognitively Grounded Model of Computational Identity

The cognitively grounded model of computational identity of the AIR Project is summarized in Fig. 3. It forms the second pillar of our approach.

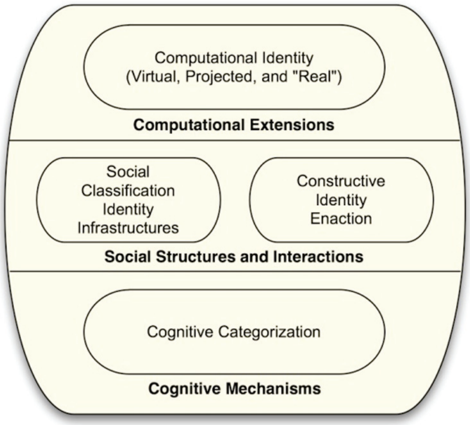

Computational Identity (Virtual, Projected, and "Real")
Computational Extensions
Social Classification Identity Infrastructures
Constructive Identity Enaction
Social Structures and Interactions
Cognitive Categorization
Cognitive Mechanisms

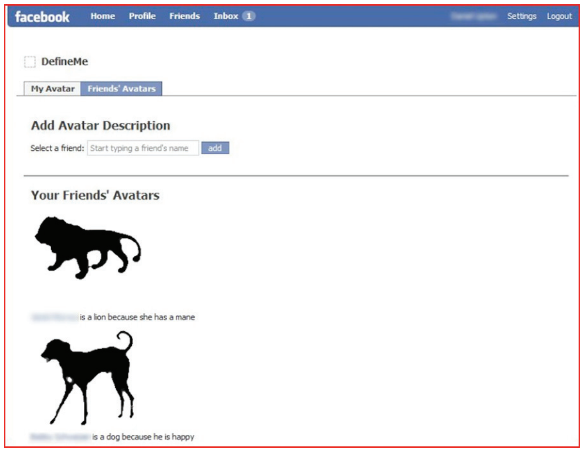
Fig. 3 A Cognitively Grounded Model of Computational Identity

Our digital identity models serve to critique infrastructures of social classification that can, unfortunately, often serve to reify naïve models of identity, and which do no capture the dynamic, constructive, and enacted identity phenomena encountered in everyday experience. This model is an analytical construct used to help to understand the interplay between underlying infrastructures, such as the technical underpinning described above, and the subjective interpretation of digital identities. The utility of the model is that we can identify where schisms exist between a technical structure (e.g., a data structure specifying a player character as a string called a "priest" and an associated "heal" procedure that allows addition to an integer called "hit points) and a real world idealized cognitive model as encoded in a classification infrastructure such as "occupation description" (e.g., the description of a priest perhaps as either someone versed in a metaphysical body of knowledge or as merely the facilitator of a particular type of institution). We can then construct new infrastructures, using techniques such as suggested in the AIR Project, that more closely align these structures and models in order to construct the hybrid of computationally afforded identities and real-world identities that James Gee calls the "projected identity" as shown in cognitively grounded AIR model (e.g., a player taking on the role of a priest in a computer role-playing game and trying to be helpful and supportive to her or his friends). The key here is that our understanding of both computational structures and the ways that users interpret them is based in imaginative cognition processes such as conceptual categorization and blending.

4. *DefineMe – Chimera:* A Critical Identity Construction Social Networking Application

The first system constructed using the AIR theoretical framework is a Facebook application entitled *DefineMe*, the first version of which is called *Chimera*. Specifically, we implement aspects of Lakoff's metonymic idealized cognitive models for categorization to allow users to co-construct their own and others' avatars as boundary objects. (Lakoff 1987) The premise behind *DefineMe* is to allow users to define each others' avatars using both commonplace and abstract metaphors. Users can append metadata to other peoples' profiles to drive dynamic generation of avatar images. The initial content domain consists of animal metaphors that can be mixed-and-matched algorithmically. Animal metaphors are potent entrenched metaphors for human personality. (Turner 1996) (e.g., sneaky weasels or docile sheep), however this animal metaphor-based version is only an initial experiment. The model extends to more directly socially engaged categories such as social scenes, fashions, or movements.

The *DefineMe* database is designed to be lightweight, dynamic, and extensible, while implementing categorical relationships between members. When comparing profiles, *DefineMe* is designed to match lexical items and logical relations directly, or it can compare the structures of profiles following insights from the analogical structure-mapping engine (SME) developed by Ken Forbus et. al. (Forbus 2001; Gentner 1983) The *DefineMe – Chimera* application reported on here focuses on creating metaphorical projections as described above. The *DefineMe* database relies on tags to create additional descriptors for each category or member. For instance, one user could describe her friend as a 'lion' (which would be the member) because she is 'strong' (which is the tag). Another user could add an additional tag, stating that she is a 'lion' because she is 'carnivorous.' These tags can comprise vertical parent-child links (e.g. a 'lion' is-an 'animal') or horizontal implicit links (e.g. in another user's profile a 'lion' is-an 'Ethiopian symbol,' yet the system may still create a category linked by the concept 'lion'). The fact that users define other users has the potential to both entertain and agitate, regardless it allows for critical inquiry into the phenomena echoing real-world labeling.



Add Avatar Description

Select a friend: Start typing a friend's name

Your Friends' Avatars

[redacted] is a lion because she has a mane

[redacted] is a dog because he is happy

Fig. 4 A screenshot of the DefineMe – Chimera facebook interface

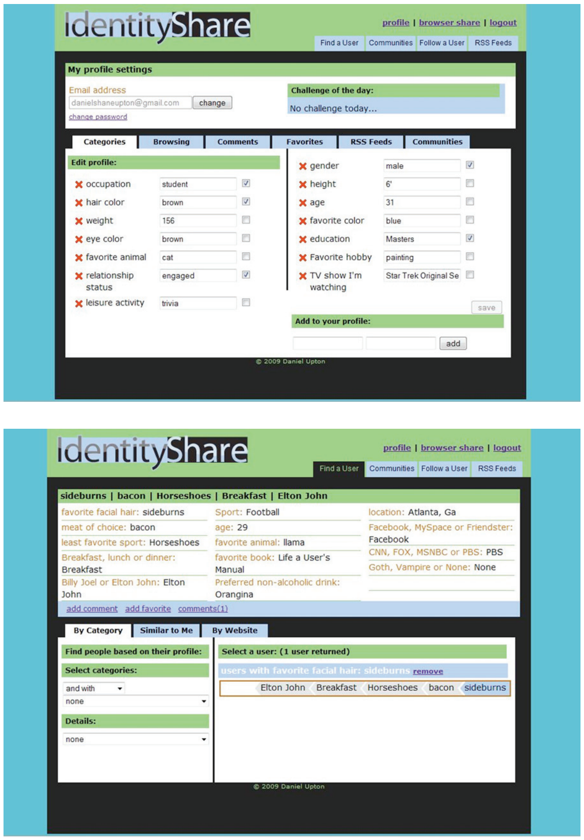
Following the work of Eleanor Rosch cited in (Lakoff 1987), the tagging system can also be used to define aspects of categories themselves. For instance, a 'robin' tag can be added to the category, 'birds,' to define the prototype of that category. In this way, members can belong to multiple groups, but individuals can represent the prototypical members of groups. In this early version, each user is seen as a member of each assigned animal category as well. This membership allows the system to use an individual as a prototype stand-in for the category. For instance, rather than just labeling a friend as a lion, one could state that your friend, Emily, is like your friend Bobby because she is brave. The system can then take all of Bobby's attributes and apply them to Emily's chimera. This relatively lightweight structure avoids some of the pre-defined categorization built into many social networking infrastructures, and has the potential to explore some of the more nuanced identity phenomena mentioned in the theoretical framework above.

5. *Identity Share:* A Critical Identity Construction Social Networking Application

Identity Share, a social networking site for "non-friends," and Daniel Upton's MS thesis project in Digital Media, is also developed under the umbrella of the AIR project. The system allows for social networking by providing users with facilities to construct profiles, follow and comment upon other users, and perform game-like tasks that encourage users to consider exploring both like and different profiles of others. *Identity Share* offers a novel means of self-representation based upon open-ended categories and tags. Standard profile models that typically include normative categories such as name, age, gender, location, and race are replaced with a customizable list that exists as a database, growing as more categories are added. Database consistency is maintained by giving users typeahead functionality when adding custom categories and by presenting existing categories in order from most common to least common. The database structure is based upon the same layout used by *DefineMe – Chimera*. Users can select which categories are most important to them by indicating that they are primary to the user using checkboxes. By allowing for primary selection of categories, we consider the system to implementing centrality phenomena from the cognitive science theory above, i.e., "the idea that some members of a category may be 'better examples' of that category than others," to a users profile. (Lakoff 1987) This means that a user's profile, as a collection of categories that define a user, is no longer viewed as just a set of static characteristics that are true about this user, but rather as a complex set of characteristics where some may be "truer" or more definitional to the user's self-conception. To take this even further, in a future implementation *Identity Share* could offer a ranking system for each category, thereby not only providing centrality, but a centrality gradience, "the idea that members (or sub-categories) which are clearly within the category boundaries may still be more or less central." (Lakoff 1987) This offers a new dynamic to social network profiling that doesn't currently exist on the popular social networking sites.

6. Ethical and Humanistic Implications

When social stakes are low, many people are inclined to reveal their baser selves. Indeed, in a project such as *DefineMe – Chimera* the potential for using the system to ridicule is quite apparent. Likewise, the ability to anonymously follow users' web usage experiences in *Identity Share* offers a potential that may seem to verge on the voyeuristic. Yet, these potentially disempowering uses are not seen as drawbacks of the systems. Each system is considered to be a culturally situated critical intervention, rather than a usability oriented productive application. In bringing to light more nuanced and imaginative identity phenomena, such as potential ostracism, prejudicial exoticizing of the other, or unflattering labeling, we hope to also provide the potential to disempower such phenomena through dialogic engagement. These systems can be considered cultural productions, or digital media art projects, in the sense that they are provocative cultural interventions situated in an environment increasingly encroached upon by hegemonically enforced, often corporate, models of user identity. As such, the systems succeed only to the degree that they engage users as evocative systems, inspire critical thought, and are construed as adequate for capturing personalities using archetypical avatars or conjure the sensation of experiencing the web through another's eyes. Beyond this, however, we see the systems as prototypes that suggest directions that could enhance the expressive and empowering potentials of productive, utilitarian, or commercial systems such as computer games and popular social networking sites with features such as self-definition of categories and deployment of imaginative metaphor.



My profile settings

Email address danielshaneupton@gmail.com

Challenge of the day:

No challenge today...

Edit profile:

occupation	student	☑	gender	male	☑
hair color	brown	☑	height	6'	☐
weight	156	☐	age	31	☐
eye color	brown	☐	favorite color	blue	☐
favorite animal	cat	☐	education	Masters	☑
relationship status	engaged	☑	Favorite hobby	painting	☐
leisure activity	trivia	☐	TV show I'm watching	Star Trek Original Se	☐

Add to your profile:

© 2009 Daniel Upton

sideburns | bacon | Horseshoes | Breakfast | Elton John

favorite facial hair: sideburns
meat of choice: bacon
least favorite sport: Horseshoes
Breakfast, lunch or dinner: Breakfast
Billy Joel or Elton John: Elton John
Sport: Football
age: 29
favorite animal: llama
favorite book: Life a User's Manual
Preferred non-alcoholic drink: Orangina
location: Atlanta, Ga
Facebook, MySpace or Friendster: Facebook
CNN, FOX, MSNBC or PBS: PBS
Goth, Vampire or None: None

Find people based on their profile:

Select categories:

and with

none

Details:

none

Select a user: (1 user returned)

users with favorite facial hair: sideburns

Elton John Breakfast Horseshoes bacon sideburns

© 2009 Daniel Upton

Fig. 5 Two screenshots of the Identity Share interface

Despite our provocative and critical interventionist stance, the systems are engineered to mitigate against abuse, and certainly distress of users is not our goal. Looking at the two systems consecutively, mitigating factors designed into the systems are as follows:

DefineMe: Chimera Design Factors

1. Users are only allowed to tag their Facebook "friends" who have added the application.
2. Users have access to a limited database of animal-types.
3. Users must "opt-in" in order to receive a generated avatar.
4. Users can "opt-out" at any time.
5. Users' database entries can be edited by moderators.
6. Users have access to only a limited format for tagging each other.
7. Users can delete entries on their profiles that others have created.

Together, we believe that these factors strongly help to avoid the system's potential to be applied in an overly negative manner. It is a contract between friends to sign up for potential compliments, teasing, and, we believe, self and social insight. Ultimately, *DefineMe – Chimera* is intended to present users with a controlled experience of torqued identity. The fractured identity of a monstrous chimerical representation is then, an accurate reflection of the limitations of applying modular and discrete classifications to a real world biography.

Regarding *Identity Share*, mitigating design factors implemented include the following:

Identity Share Design Factors

1. Users can create their own self-classifications.
2. Users can select which classifications are important to them.
3. Users can avoid or utilize normative categories such as gender or occupation.
4. Users can allow or disallow the system's tracking of their web visitation paths at will.
5. Users' real world identities are kept anonymous.

6. Users' perceived affordances to communicate with one another are highly restricted.

7. Users have full control to delete any of their data in the system.

These factors were developed over the course of iterative refinement of the project based on informal user feedback (mainly via open-ended interviews). The greatest challenge with the system was to allow for user generated categories while also pruning sparsely used and idiosyncratic database elements. A second challenge regarding anonymity and privacy is addressed by careful controls such as articulated above, and by providing quite clear and prominent information on the nature of the site. Quite contrary to being a site to allow people to "spy" on others, it is an "opt-in" site oriented toward users with a desire to share their personal styles, definitions, and web behaviors with one another. Finally, it is a system that is proposed as a balance between the limited and discrete, yet highly modular and structured, information structures provided by digital media and the continuous and transient, yet not computationally amenable, phenomena of identity as shared in the real world.

7. Conclusion

The AIR Project examines the humanistic implications of emerging technologies by seriously considering the cultural effects of user identity within current digital media and the shared sociocognitive foundations that ground their construction. Following various accounts describing the procedural nature of the computational medium (Manovich 2001; Murray 1997), the AIR Project looks at the underlying data structures and algorithms and how they implement cultural identity effects, and posits a technical framework for more deeply engaging identity semantics of classification and categorization. Technologies for implementing socially empowering or expressively critical and transformative experiences are necessary to create experiences to engage real identity social phenomena that lie at the center of so many of our political debates and rich critical fictions.

Acknowledgements

We would like to acknowledge early assistance from the undergraduate researcher Richard Shemaka. Additionally, we thank Carl DiSalvo for significant input on *Identity Share* as Daniel Upton's coadvisor with Fox Harrell.

References

Bowker, G. C. & Star, S. L. (1999). *Sorting Things Out: Classification and Its Consequences*, MIT Press, Cambridge, MA.

Fauconnier, G. (1985). *Mental Spaces: Aspects of Meaning Construction in Natural Language*, MIT Press/Bradford Books, Cambridge.

---- (2006). 'Rethinking Metaphor', in *Cambridge Handbook of Metaphor and Thought*, ed. R. Gibbs, Cambridge University Press, Cambridge, U.K.

Forbus, K. D. (2001). 'Exploring Analogy in the Large', in *The Analogical Mind: Perspectives from Cognitive Science*, MIT Press, Cambridge, MA.

Gee, J. P. (2003). *What Video Games Have To Teach Us About Learning and Literacy*, Palgrave Macmillan, New York City.

Gentner, D. (1983). 'Structure-mapping: A theoretical framework for analogy', *Cognitive Science*, vol. 7, no. 2, pp. 155-70.

Harrell, D. F. (2007). Theory and Technology for Computational Narrative: An Approach to Generative and Interactive Narrative with Bases in Algebraic Semiotics and Cognitive Linguistics, Dissertation thesis, University of California, San Diego.

---- (2008a). 'Algebra of Identity: Skin of Wind, Skin of Streams, Skin of Shadows, Skin of Vapor', in *Critical Digital Studies*, eds A. Kroker & M. Kroker, University of Toronto Press, Toronto.

----(2008b). Digital Metaphors for Phantom Selves: Computation, Mathematics, and Identity in Speculative and Fantastic Fiction and Gaming paper presented to The 29th International Conference on the Fantastic in the Arts, Orlando, FL.

Hutchins, E. (1996). *Cognition in the Wild*, MIT Press, Cambridge, MA.

---- (2000). Distributed Cognition, paper presented to International Encyclopedia of the Social & Behavioral Sciences, University of California, San Diego.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press, Chicago, IL.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western*

Thought, MIT Press, Cambridge, MA.

Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate periferal participation*, Cambridge University Press, Cambridge, U.K.

Manovich, L. (2001). *The Language of New Media*, MIT Press, Cambridge, MA.

Murray, J. H. (1997). *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*, Free Press, New York.

Santa Ana, O. (2002). *Brown Tide Rising: Metaphors of Latinos in Contemporary American Public Discourse*, University of Texas Press, Austin.

Turner, M. (1996). *The Literary Mind: The Origins of Thought and Language*, Oxford UP, New York; Oxford.

Text analysis of large corpora using High Throughput Computing

Mark Hedges
King's College London
mark.c.hedges@googlemail.com

Tobias Blanke
King's College London
tobias.blanke@kcl.ac.uk

Gerhard Brey
King's College London
gerhard.brey@kcl.ac.uk

Richard Palmer
King's College London
richard.d.palmer@kcl.ac.uk

The recent initiative of the US NEH on supercomputing [1] is just one sign that there is a growing interest in the use of highly parallel processing in the humanities. This comes as no surprise if one considers that all over the world, governments and funding bodies are investing heavily in digitization of cultural heritage and humanities research resources. At the same time, the 21st century sciences are demanding an infrastructure to support their advanced computational needs; the computational infrastructure to distribute and process the results from the Large Hadron Collider is just one example. Humanities researchers have therefore begun to investigate these infrastructures to find out whether they can be used to help analyse the extensive, newly available online resources.

We offer an example of such an infrastructure based on High Throughput Computing (HTC). HTC differs from High Performance Computing (HPC), in that the latter relies on hardware specifically designed with performance in mind, whereas the former typically uses multiple instances of more standard computers to accomplish a single computational task. The de facto standard HTC implementation is the Condor Toolkit, developed by the University of Wisconsin-Madison (http://www.cs.wisc.edu/condor). Condor can integrate both dedicated computational clusters and standard desktop machines into one computational resource.

We will present the work of the UK HiTHeR project (http://hither.cerch.kcl.ac.uk/), which created a prototype of such an infrastructure to demonstrate the utility of HTC methods to textual scholars, and indeed to humanities researchers in general. It did this by using Condor to set up a Campus Grid, which may be defined as environments which utilise existing computational and data resources within a university or other research institutions. The project used the resuting infrastructure to produce a case study based on a high profile digitisation project in the UK, addressing questions of textual analysis that would not be feasible without this infrastructure. At the same time, we will show how the application can be integrated into the web publication of text-based resources.

After demonstrating the power of HTC for standard text processing in the digital humanities, the main aim of the project is to show how digital humanities centres can be served by implementing their own local research infrastructure, which they can relatively easily build using existing resources like standard desktop networks. There have been many experiments in the digital humanities using dedicated HPC facilities, less on the application of these relatively light-weight computational infrastructures. We demonstrate the feasibility of such infrastructures and evaluate their utility for the particular task of textual analysis.

Only with such local infrastructures will it be possible to fulfil the often expressed demand of textual studies researchers to be able to experiment with the statistical methods of textual analysis rather than to be simply confronted with the results. Faced with the opportunities of HPC and HTC, these researchers frequently express the desire to transform the underlying statistical algorithms 'interactively' by changing parameters and constraints, and in this way to follow their particular interests by experimenting with the outcome of the analysis and thus gaining better insights into the structures of the text. If the humanities researcher has go to dedicated supercomputing centres, such an approach is more difficult to maintain, as it will depend on the relationship with that supercomputer centre. HiTHeR has thus two research goals: (1) to carry out textual analysis in a parallel computing environment and (2) to investigate new types of e-Infrastructures for supporting the work of digital humanities centres.

HiTHeR infrastructure - Case Study

Automatic textual processing is relatively well researched and can rely on a large range of specialised algorithms and data structures to process textual data. In the digital humanities, many of these algorithms have been tried on complex historical or linguistic collections. Language modelling, vector space analysis, support vector machines or LSI are only some of the machine learning approaches that have attracted growing interest in the digital humanities. In the HiTHeR project, we focused on relatively simple processing, which nevertheless has proven to be highly useful to many humanities research institutions. A recent study in the ICT needs of humanities researchers [2] found out that text analysis tools and services are still generating most interest among humanities researchers. Among these text analysis tools, the comparison of 2 or more texts was seen as the most useful tools. Such tools help with many textual studies activities from the comparison of different versions of texts to finding texts about the same topic in large textual collections. Such comparisons may rely on stable algorithms but are often costly in terms of the computational resources needed, as each document in a collection needs to be compared to all other documents in the collection. Also, the digital textual resources processed are often 'dirty', containing a high proportion of transcription errors, because of the problems of digitising older, irregular print. This leads to further increases in computational size and complexity, as more advanced methods are needed to reduce the "noise" from the OCR processes. Overcoming the complexities of machine based learning in the humanities, was therefore recognized quite early as a use case for an advanced computational infrastructure.

The corpus used for the HiTHeR case study is the Nineteenth Century Serials Edition (http://www.ucse.ac.uk/), which contains circa 430,000 articles that originally appeared in roughly 3,500 issues of six 19th Century periodicals. Published over a span of 84 years, materials within the corpus exist in numbered editions, and include supplements, wrapper materials and visual elements. Currently, the corpus is explored by means of a keyword classification, derived by a combination of manual and automated techniques. A key challenge in creating a digital system for managing such a corpus is to develop appropriate and innovative tools that will assist scholars in finding materials that support their research, while at the same time stimulating and enabling innovative approaches to the material. One goal would be to create a "semantic view" that would allow users of the resource to find information more intuitively. However, the advanced automated methods that could help to create such a semantic view require greater processing power that is available in standard environments. Prior to the current case study, we were using a simple document similarity index to allow journals of similar contents to be represented next to each other. The program used the lingpipe (http://alias-i.com/lingpipe/) software to calculate similarity measures for articles based on frequencies of character n-grams within the corpus. A character n-gram is any subsequence of n well defined characters. Initial benchmarks on a stand-alone server allowed us to conclude that, assuming the test set was representative, a complete set of comparisons for the corpus would take more than 1,000 years. Consequently, we ran a sequence of systematic experiments, carrying out different text analysis of the selected corpus, to provide benchmarks for the throughput improvements provided by the grid environment. The detailed results will be presented in the presentation.

In the experiments, we have set up an institutional CampusGrid using Condor at King's College London on spare servers and desktops (in use during the day) within 2 departments. No new hardware had to be bought. We than ran several text mining algorithms on a subset of the data (the "English Women's Journal"—which has the highest OCR quality) which have been adapted locally so that parts of the code can be run in parallel. This has reduced the processing time from a few days to a few hours.

To conclude: One driver of the project is the NCSE corpus, for which the project addresses a genuine research need to be able to create new semantic views on textual resources automatically. But, more widely than this, we see the project as an opportunity to start building the e-infrastructure required to support humanities research that has complex (or simply large) computational requirements.

References

[1] National Endowment for the Humanities (NEH) Digital Humanities Initiative (Workshop on Supercomputing & the Humanities (July 11, 2007), http://www.neh.gov/whoweare/cio/odhfiles/HHPC.Workshop.07.11.2007.final.pdf

[2] Toms, Elaine and O'Brien, Heather L.: Understanding the information and communication technology needs of the e-humanist, Journal of Documentation, vol 64, 2008.

MAPS: Manuscript map Annotation and Presentation System

Linking formal ontologies with social tagging to (re-)construct relationships between manuscript maps and contextual documents

Charles van den Heuvel
Virtual Knowledge Studio for Humanities and Social Sciences
charles.vandenheuvel@vks.knaw.nl

Introduction

The proposed paper describes a pilot entitled MAPS of the Virtual Knowledge Studio for the Humanities and Social Sciences (VKS) of the Royal Netherlands Academy of Arts and Sciences (KNAW), the Dutch National Archives, Leiden University Library and the Department of Information and Computing Sciences of Utrecht University intended to disclose manuscript map collections in the Netherlands and potentially abroad for individual and institutional research in the humanities. The research is embedded in the Spatial/GIS lab of a digital humanities collaboration of institutes of the Royal Academy, the so-called "Alfa-Lab" that started in March 2009.

It builds upon research and existing "draw over image" software, developed within the context of the Paper and Virtual Cities project, a collaboration between the University Groningen and the Virtual Knowledge Studio. MAPS provides a system in which users reconstruct historical contexts by means of annotations and bottom-up geo-references, here called "spatial tags" complementary to the formal ontology based Encoded Archival Description standard. In a technological sense this implies the interoperability between (semi-)structured and "fuzzy" annotations in the form of tagged draw-over-images and the (semi-)automatic linking of maps with contextual documents. Besides technical issues, methodological questions are discussed regarding the role and authority of annotations and tags of professional and non-professional humanities researchers in a Web 2.0 environment. The requirements for use of manuscript maps in humanities research are mapped by means of observational user studies of annotation practices and by testing interfaces.

Manuscript Maps

After the Fall of Antwerp (1585), Amsterdam, with famous map printers/ publishers such as Blaeu, became in the 17th Century a leading cartographical center in the world. Research of this predominant position in commercial cartography is relevant for historians of cartography, for historians of the book and publishing and for cultural studies, such as art history. However, this emphasis on printed commercial cartography also blurs the importance of manuscript maps for research in a much wider field of individual historians or of cultural heritage, planning and policy institutions that use historical maps for technical or administrative purposes. Manuscript maps are significant not only for their quantity, the Dutch National Archives alone count approximately 300.000 manuscript maps (including drawings) from Dutch administrative bodies, but also for their quality as unique historical sources. (De Vries 1989, Zandvliet 1998, Heuvel 2003 and 2004). Unfortunately, manuscript maps are less accessible; their proper meaning can only be understood when they are re-linked to the documents they originally were intended to illustrate. Due to 19th and 20th century archival processes many maps lost their contexts when they were separated from the documents to which they belonged. Encoded Archival Description (EAD) and its extension Encoded Archival Context (EAC), used as standards in many Dutch archives, libraries and museums (including the Dutch National Archives), are powerful description and contextualization formats to encode links between maps and documents dispersed over cultural heritage institutions. Although the first experiments (coordinated by the Dutch National Archives for the Netherlands and with Leiden University Library as participant) with such formats are promising and will be fully exploited, this linking process to reconstruct original contexts and to create new ones is time-consuming and cannot be done by cultural heritage institutions on their own. A more productive way is to involve so many users as possible in the (re) establishment of these historical relationships by enriching these manuscript maps with annotations, and to link these to archival documents as described in inventories.

Annotation

In order to identify maps and to use them for historical research, it is important to link them to geographical space. Geo-referencing allows combinations and overlays of maps with minimum distortion. However, this process is quite difficult for manuscript maps that vary so much in scale, precision, color etc. Instead of complex geo-referencing systems we opt for existing draw-over-image software, developed within the context of the Paper and Virtual Cities project, a collaboration between the University Groningen and the Virtual Knowledge Studio that allows us to select any object on the map for annotation in XML. (Heuvel & Koster 2007, Benavides& Heuvel 2008), On the one hand the map annotator module will be used by individual experts in a conventional way to add knowledge through structured annotations based on professional research. Leiden University Library is a research library, whose special collections are used for the greater part by scholars. On the other hand much more fuzzy and fragmented knowledge is available from users, who work more incidentally with maps, e.g. genealogists, local historians and private collectors. Their input through Web 2.0-techniques may be of great help to reconstitute the lost context of historical maps. They provide the required critical mass; 85% of the visitors to the Dutch National Archives are genealogists. However, we have to learn from Web 2.0 failures (Crotty 2008) and avoid mismatches between tools and the researchers that use them. Therefore analysis of annotation practices and usability testing of interfaces with focus groups of researchers in the humanities makes part of the project. The methods used by E. Heere (2008) for analyzing the use of a historical GIS interface, will be extended to the analysis of practices of annotating/tagging by researchers in a Web 2.0 environment. Another issue that will be addressed is the authority of annotations by expert and lay-experts. Hereto we try to build on existing software (such and GENTECH and HarvANA) that allow expressing and visualizing the provenance and levels of authority of certain assertions. (Hunter 2008)

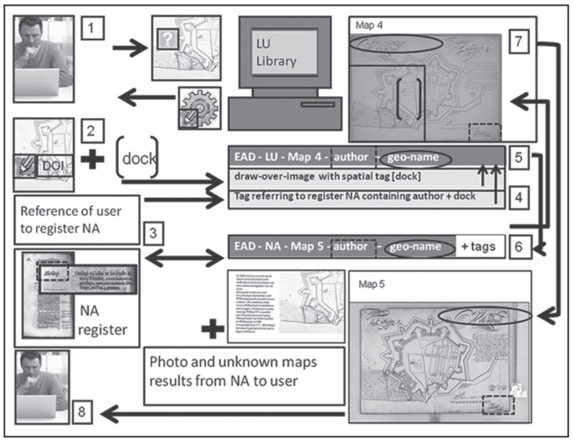
Presentation

Users have an easy access to the archive inventories described in EAD in order to search and annotate digitized manuscript maps and related sources. Annotations of geographical space and time are described in specific XML vocabularies, such as Geographical Markup Language and Historical Event Markup Language. After linking these clusters of knowledge have to be made available to researchers according to their needs. This is done in mash-ups maps (Google Maps, using the KML standard). Because map annotation will be an ongoing process, the user interface should be integrated in a coherent system offering not only reference lists (tag lists, gazetteer and archive inventories), but also the already tagged documents.

System

MAPS enhances the research infrastructure by building an annotation infrastructure for user generated research content in the humanities, by geo-semantic data integration (in this case maps, annotations and archival data), and by the presentation in suitable user-interfaces for researchers in the form of mash-ups. (Ennals 2007) The geo-semantic data integration between maps, annotations and archival data will be based on (semi-) automatic linking, using multi-agent technology, between profiles of (historical) archive creators, typologies of maps and profiles of individual experts and lay expert researchers and research groups and cultural heritage, planning and policy institutions that use historical maps for technical or administrative purposes.

Scenario


1
?
LU
Library
Map 4
7
2
DOI
dock
EAD - LU - Map 4 - author - geo-name
draw-over-image with spatial tag [dock]
Tag referring to register NA containing author + dock
5
4
Reference of user
to register NA
3
EAD - NA - Map 5 - author - geo-name + tags
6
NA
register
Map 5
Photo and unknown maps
results from NA to user
8

Scheme of links between user and cultural heritage institutions in Manuscript Map Annotation and Presentation System - MAPS

The Dutch researcher Peter Sigmond studies Dutch habours between 1500-1800 for his PhD research. In chapter of his dissertation with the title *Nederlandse Zeehavens tussen 1500-1800* he discusses a report and some designs of an engineer Alberdingh of 1688 for the creation of dock harbors in several Dutch harbors as alternative to the marine one in Amsterdam that gradually silted up and was not accessible anymore for heavy warships. Sigmond ordered a photograph of a part of a manuscript map number 4 of one of these designs for the town of Willemstad in the collection of Leiden University Library (LU). In our scenario this library points him to the draw-over-image tools [1] that allows him to indicate the exact contours of the object on the digital map in its image repository, that he wants to use for his illustration. He sends this image with the tag dock to LU [2]. His PhD also contains a reference to a report in the Dutch National Archives (NA) that describes the pro and cons of the various harbor cities for which the designs were made. He tags this information to the draw-over-image (DOI) and sends it to LU [3]. DOI and tag are linked to the metadata description of the map 4 in EAD that has already a link to three others files describing maps belonging to the same project [4]. Numbering of the maps also make clear that there are some missing maps. Since the report refers to the Dutch National Archives a link is made to sources described with EAD metadata of this archive, including the DOI and tags [5] and reconnected with the maps of LU [6]. In a query on expected archive creator profile, in this case military, the database shows that there is another design of the same year for the same project map 5 and another later copy of it. The image of LU and the metadata of all their relevant images are combined with those of NA[7] and send with the photographical order to the user [8]. With this information the series of designs in Leiden University Library is completed with one extra drawing from the NA. The PhD researcher found out that there were extra drawings. Finally the National Archive found that there were more drawings belonging to the report in their collection.

References

Benavides J. and Heuvel, C. van den (2008). The function and accuracy of old Dutch urban designs and maps. A computer assisted analysis of the extension of Leiden (1611) in *Digital Humanities 2008 Conference*, University of Oulu, Finland June 24-29, 55-56.

Crotty, D. (2008). *Why Web 2.0 is failing in biology.* Harbor Protocols Cold Spring http://www.cshblogs.org/cshprotocols/2008/02/14/why-web-20-is-failing-in-biology/

EAD/EAC. http://www.iath.virginia.edu/eac.

ECAI. http://www.ecai.org/

Ennals, R. and Gay, D. (2007). User-friendly functional programming for web mashups. In *Proceedings of the 2007 ACM SIGPLAN international Conference on Functional Programming* (Freiburg, Germany, October 01 - 03, 2007). ICFP '07. ACM, New York, NY, 223-234.

GENTECH: http://www.ngsgenealogy.org/ngsgentech/projects/Gdm/Gdm.cfm.

Heere, E. (2008). *GIS voor historisch landschapsonderzoek: opzet en gebruik van een historisch GIS voor prekadastrale kaarten,* PhD thesis Utrecht University 27 June 2008.

Heuvel, C. van den (2003) "Atlasticity", Problems in defining and the digitizing military manuscript atlases of the Low Countries. In [ed. I. Warmoes, E. d'Orgeix, C. van den Heuvel] *Atlas Militaires Manuscrits Européens (XVIe-XVIIIe siècles). Forme, contenu, contexte de réalisation et vocations*, Actes des 4es journées d' étude du Musée des Plans-reliefs, Paris (18-19 April 2002) Paris pp. 11-26.

Heuvel, C. van den (2004). Mapping Mixed Maps. Historical and future constructions of time and space in urban cartography. In *Le temps 129e congrès des sociétés* *historiques et scientifiques, Colloque IV Le temps des cartes. Monde des cartes, Bulletin du Comité français de cartographie [Numéro special]*,pp. 23-40.

Heuvel, C. van den, Koster E. (2007). Tussen papieren en digitale kaarten. De annotatie van een atlas en bijbehorende kaarten van Woerden. In *Aangeraakt. Boeken in contact met hun lezers. Een bundel opstellen voor Wim Gerritsen en Paul Hoftijzer,* red. Kasper van Ommen, Arnoud Vrolijk, Geert Warnar, *Kleine publicaties van de Leidse Universiteitsbibliotheek Nr. 75*, pp. 150-156.

Hunter, J., Khan, I., and Gerber, A. (2008) HarvANA -Harvesting Community Tags to Enrich Collection Metadata. In ACM IEEE Joint Conference on Digital Libraries, JCDL June 16 - 20, 2008. Pittsburgh.

Vries, D. de [ed.] (1989). *Kaarten met geschiedenis 1550-1800. Een selectie van oude getekende kaarten van Nederland uit de Collectie Bodel Nijenhuis* HES, Utrecht

Zandvliet, K. (1998). *Mapping for Money. Maps, plans and topographic paintings and their role in Dutch overseas expansion during the 16th and 17th centuries*, Batavian Lion international, Amsterdam.

An Exercise in Non-Ideal Authorship Attribution

David L. Hoover
New York University
david.hoover@nyu.edu

Much has been written about the appropriate conditions for non-traditional authorship (e.g., Rudman 1997). Substantial known samples by the possible authors of a disputed text should be available, they should be controlled for genre, point of view, etc., and additional similar texts should be available from other contemporary authors to show that false positive identifications are unlikely. Less attention has been paid to authorship problems compelling enough to demand an attempt in spite of less-than-ideal conditions. One such case is Maria Ward's Female Life Among the Mormons (1855), one of four widely read books published in 1855-56 that were largely responsible for the litany of anti-Mormon literature of the late 19th and early 20th centuries, including works by Arthur Conan Doyle, Zane Grey, Robert Louis Stevenson, and Mark Twain. (The first edition of Female Life sold at least 40,000 copies; within three years it had been translated into Danish, French, German, Hungarian, and Swedish (Worldcat).)

Female Life, published as an anonymous autobiographical account, was early panned and labeled fiction (*New York Times* 1855), and there is strong evidence that 'Maria Ward,' who also claims to have 'edited' Austin Ward's *The Husband in Utah* (1857) is a pseudonym (the book was later published as *Male Life Among the Mormons*; I refer to it as *Male Life* below). Confusingly both Ward books are sometimes categorized as fiction, and Mrs B. G. Ferris is sometimes listed as a (second) author of *Female Life* (Worldcat). Moreover, both Mr and Mrs Ferris wrote non-fictional anti-Mormon books–his a history of Utah and Mormonism, *Utah and the Mormons* (1854), hers a memoir, *The Mormons at Home* (1856). Finally, efforts to confirm Ferris's authorship of *Female Life* led Arrington and Haupt (1968: 253) to conclude that she did not write it, but that her book may have provided the basis for *Male Life*.

A Computational Analysis

The case of *Female Life* certainly qualifies as a non-ideal authorship attribution problem. Each of the four authors of interest apparently wrote only one book. All four books are framed as nonfiction, but some extraordinary errors and the fact that 'Maria Ward' and 'Austin Ward' are likely pseudonyms suggests that their books are fiction, makes their gender doubtful, and suggests that they may be two pseudonyms for one person. Additionally, only the middle third of Mrs Ferris's book deals with the Mormons, and the attribution of *Female Life* to her is tenuous and unexplained. The most important facts are these:

B. G. Ferris	*Utah and the Mormons*	Non-fiction	3rd-Person	Male
Mrs B. G. Ferris	*The Mormons at Home*	Non-fiction	1st-Person	Female
Austin Ward (Maria Ward?)	*Male Life*	Fiction?	1st-Person	Male?
Maria Ward (Mrs Ferris?)	*Female Life*	Fiction?	1st-Person	Female?

Given this less-than-ideal scenario, can we shed any light on the authorship of *Female Life*? I begin with a group of large, Mormon-related, first-person fiction and nonfiction texts. With many candidate authors, any similarities between Mrs Ferris and Maria Ward will be more significant. In Fig. 1, only Kane and Stenhouse are represented by pairs of texts, and Stenhouse's texts fail to cluster. Yet 'StenhouseMa' is an excerpt from *Tell it All* that supposedly relates Mary Burton's story in her own words, so that this is not a definite error. Adding Mrs Ferris and Maria Ward (Fig. 2), does not support Ferris's authorship of *Female Life*: the two authors' texts are widely separated, but the close clustering of the Wards' texts suggests a single author. (In the captions, NP 70% indicates that no personal pronouns are included and words are eliminated if a single text accounts for more than 70% of their occurrences. Pronouns are closely tied to the number and gender of characters; words that are frequent because of a single text are typically character names.)

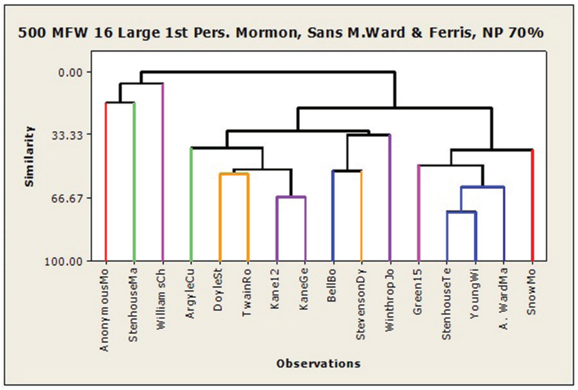


Fig. 1 Large Mormon 1st person only–without Ward & Ferris

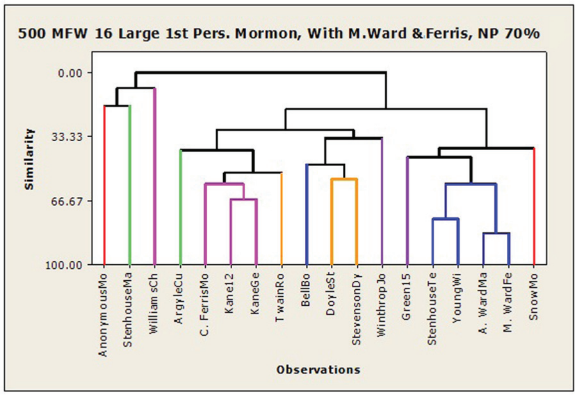


Fig. 2 Large Mormon 1st person only–including Ward & Ferris

Large 1 and 3 Person Mormon and Non-Mormon Texts

A larger set of texts with more pairs by single authors allows me to train the method on similar texts, select the most effective analysis, then insert Ward and Ferris and re-test–compensating for the lack of primary author training texts. Consider then 32 large Mormon and non-Mormon texts (Fig. 3). Eleven of thirteen pairs by known authors cluster together, capturing known authorship very effectively, though most contain Mormon non-Mormon texts. The two failures may not be errors: Stenhouse's texts are discussed above, and Bell claims only to have 'prepared' Stephens's *Rebel Cousins*. Adding Maria Ward's and the Ferrises's texts (Mrs Ferris's in three sections, isolating the Mormon part), produces Fig. 4. The known texts cluster as before, and the Wards form a tight cluster separate from the Ferrises, whose own close clustering may suggest collaboration. (A smaller set of Mormon-related texts with the dialogue deleted to eliminate any effects of different proportions of dialogue and narration, produces similar results.)

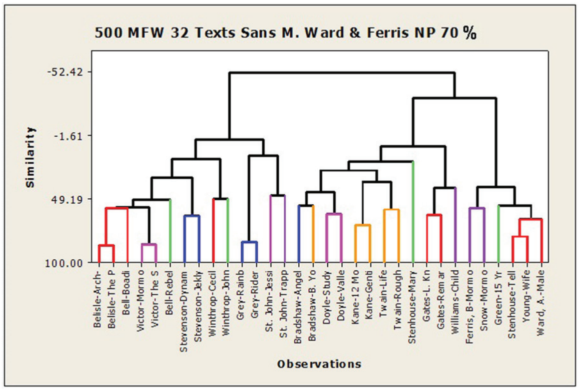


Fig. 3 Large 1 and 3 person Mormon and non-Mormon texts–without Ward & Ferris

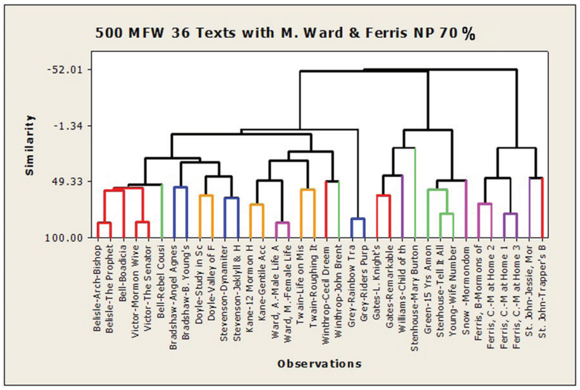


Fig. 4 Large 1 and 3 person Mormon and non-Mormon texts–including Ward & Ferris

Delta Analysis

Delta analysis is problematic here because of the complexity of the case, but several analyses with different configurations of primary and secondary authors never strongly identify Mrs Ferris as the author of *Female Life,* though a few identify her as the author of *Male Life* when Maria Ward is not among the primary authors. When either of the Wards is included among the primary authors, however, he or she is regularly and strongly identified as the author of the other Ward's text. (For discussion of Delta, see Burrows 2002 and Hoover 2004.)

T-Tests

It seems worthwhile to try another approach using t-tests. Each test assumes that two of our three main authors (Maria and Austin Ward and Mrs Ferris) are different, uses t-tests to create a set of words that strongly differentiate them, and then uses these words to test the third author with PCA (principal components analysis). (The method is inspired by Burrows 1992.) This makes sense, however, only if we know what to expect when each set of assumptions is false, so I begin with several sets of three known texts by one, two, or three authors.

In two of three tests involving three texts by one author, the third text intermingles with one of the two texts distinguished by the t-tested words. In the third test, however, the three texts remain quite distinct. Thus, if Mrs Ferris wrote all three texts and they cluster separately, no conclusions can be drawn. With three texts by two authors, t-tests that distinguish two texts by one author and those that distinguish two texts by two authors produce very different results. In the former case, all three texts typically cluster separately (Fig. 5); in the latter, the two texts by the same author are separate from that of the third on component one, but not from each other. In two of the three cases tested, they also mingle on component two (Fig. 6). T-tests involving two texts by one author also typically produce far fewer discriminating words (275) than those involving two authors (677). Finally, in all tests involving three authors, the three texts remain quite distinct, whichever authors are assumed to be different.

Once the three scenarios have been explored, we can compare these results with tests on our authors of interest. Assuming the two Wards are different and testing Mrs Ferris as an unknown (narrative only, approximately 27,000 words for each of the Wards and 18,000 for Ferris) yields Fig. 7. (Tests on the three whole texts produce similar results.) T-tests involving one author, three authors, or two authors (when the wrong assumption of difference is made) can produce similar graphs, though the small number of discriminating words supports the last possibility. But Fig. 8. and Fig. 9, with one of the Wards tested against Ferris and the other Ward as unknown, produce patterns like that in Fig. 6, patterns consistent with the results from Delta and the various cluster analyses above: Mrs Ferris is unlikely to have written either *Female Life* or *Male Life*, and Maria and Austin Ward seem very likely to be two pseudonyms for the same person (whose identity remains a mystery).

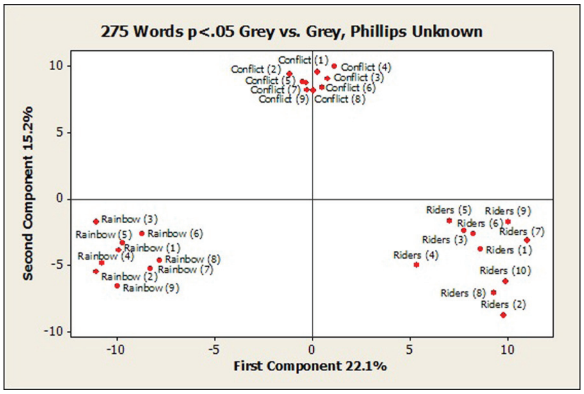


Fig. 5 Grey (Rainbow) vs Grey (Riders), with Phillips unknown (275 words p<.05)

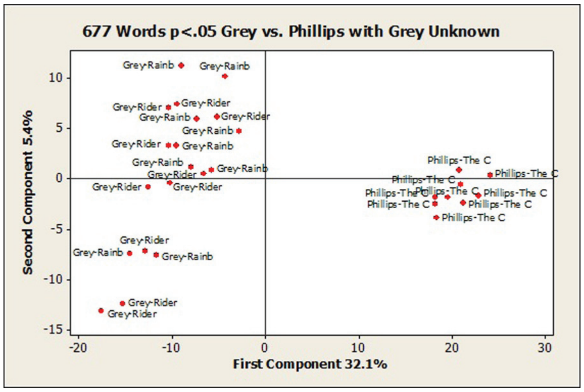


Fig. 6 Grey (Rainbow) vs. Phillips (Conflict) with Grey unknown (677 words p<.05)

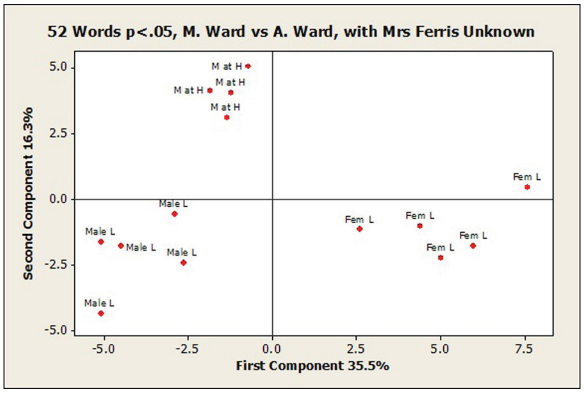


Fig. 7 Austin Ward vs Maria Ward, with Mrs Ferris unknown

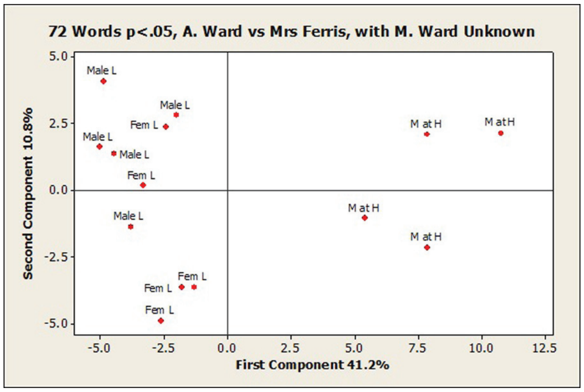


Fig. 8 Austin Ward vs Mrs Ferris, with Maria Ward unknown

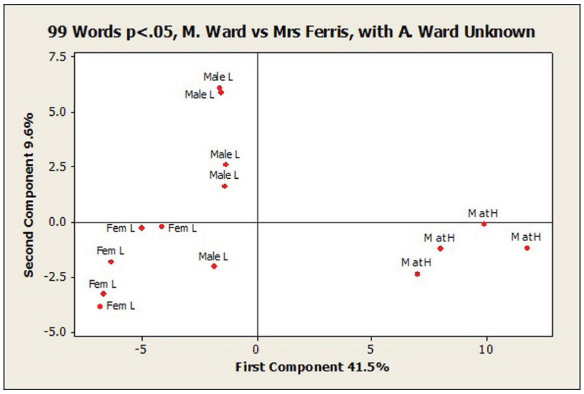


Fig. 9 Maria Ward vs Mrs Ferris, With Austin Ward Unknown

Conclusion

Considering the level of difficulty of this authorship problem, the results seems quite persuasive. They suggest that combining several different approaches across many analyses can help to compensate for the lack of training texts. Finally, comparing results based on false and true assumptions in simulations with known authors to the results of similar tests involving the texts in question can provide worthwhile results even under conditions that are far from ideal.

References

Arrington, Leonard J., and Jon Haupt. (1968). Intolerable Zion: The Image of Mormonism in Nineteenth Century American Literature, *Western Humanities Review*, 22: 243-260.

Burrows, John F. (1992). Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information, *LLC* 7: 91-109.

Burrows, John F. (2002a). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship, *LLC* 17: 267-287.

Ferris, Benjamin G. (1854). *Utah and the Mormons*. New York: Harper & Brothers.

Ferris, Mrs Benjamin G. (1856). *The Mormons at Home*. New York, Dix & Edwards.

Hoover, David L. (2004). Testing Burrows's Delta, *LLC* 19: 453-475.

Review of *Female Life Among the Mormons*. (1855). *The New York Times*. July 14: 3.

Rudman, Joseph. (1997). The State of Authorship Attribution Studies: Some Problems and Solutions, *Computers and the Humanities* 31: 351-65.

Ward, Austin N. (1857). *The Husband in Utah*. Maria Ward (ed.). New York: Derby & Jackson; republished as *Male Life Among the Mormons,* Philadelphia: John E. Potter and Company, 1863.

Ward, Maria. (1855). *Female Life Among the Mormons*. New York: J. C. Derby.

Worldcat. (2001-2008). Dublin, Ohio: OCLC Online Computer Library Center, Inc. Online.

Modes of Composition in Henry James: Dictation, Style, and *What Maisie Knew*

David L. Hoover
New York University
david.hoover@nyu.edu

> A hundred things hummed at the back of her head, but two of these were simple enough. Mrs. Beale was by the way, after all, just her stepmother and her relative. She was just—and partly for that very reason—Sir Claude's greatest intimate ('lady-intimate' was Maisie's term) so that what together they were on Mrs. Wix's prescription to give up and break short off with was for one of them his particular favourite and for the other her father's wife.
> Henry James, *What Maisie Knew* (1897)

The difference between the styles of Henry James's early and later novels is well-known. No author's style is likely to remain constant over a time-span as long as that between James's first novel in 1871 and his last complete novel in 1911, but the change in James's style is so extreme that many readers and critics consider the later novels too obscure and enjoy only the early novels. Others champion the later novels and consider the early novels immature. One critic notes that James's late 'distortions' often 'obliterate the normal elements of connection and cohesion. When he has undone the usual ties, his meanings float untethered, grammatically speaking, like particles in colloidal suspension' (Short 1946: 73-4). Although most attention has been paid to syntax, changes in James's vocabulary are equally interesting, and the frequencies of the most frequent words alone very effectively identify when each novel was written (Hoover 2007).

One often-mentioned reason for the change in James's style seems important from a literary perspective and is especially amenable to and appropriate for computational investigation: In 1896, while writing *What Maisie Knew,* James began dictating because of persistent pain in his wrist. His most famous typist believed that 'The different note [of the later novels] was possibly due more to the substitution of dictation for pen and ink than to any profound change of heart' (Bosanquet 1924: 254). Edel's biography of James links the elaborate and convoluted late sentences with the change in mode of composition and reports that some of James's friends claimed to be able to tell which chapter in *Maisie* was the first one he dictated (1969: 176-77). Recent interest in media and technology, especially at the turn of the twentieth century, has given new significance to the claim (Cappello 2007; Thurschwell 2001; Seltzer 1992). The fact that it is mentioned by Marshall McLuhan (McLuhan & Zingrone 1995), and that it appears in *Barron's Notebooks* (a study guide for students) (1986), the online guide to the *Dragon Naturally Speaking* dictation-to-text program (Newman 2000), an article on voice recognition in composition (Honeycutt 2003), *Wikipedia* (2008), *The Ivanhoe Game* (Bethany 2002), and the fiction of Cynthia Ozick (2008), shows its ubiquity, appeal, and relevance to digital media and digital humanities.

The fact that James's later style is generally dated to the late 1890s, just when he begins dictating, has undoubtedly encouraged this idea. Despite of the widespread currency of the belief that dictation significantly affected James's style, I know of no attempt to present any solid evidence for it. Yet there is strong evidence for the gradual and unidirectional development of James's style over his career that casts doubt on any significant effect of the change to dictation. Consider the extraordinary pattern of adoption of characteristically late words and abandonment of characteristically early words over time shown in Fig. 1. To produce this figure, I created combined word frequency lists for the seven earliest and eight latest novels, then selected all words that are at least three times as frequent in the early novels as in the late and vice versa. I calculated the percentage of tokens represented by early and late words in a selection of novels and novellas throughout James's career. Note how regularly the frequency of early words declines and late words increases. These trends continue through the middle period and include novels and novellas that had no part in the creation of the word lists. The later style is associated more with syntax than with vocabulary, but the regularity of the vocabulary changes certainly offers no support for an abrupt shift between the handwritten novels that precede and the dictated novels that follow *Maisie*.

Now let us examine *Maisie* itself for evidence of a significant local shift in style caused by the switch to dictation. Evidence from the most frequent words is shown in Fig. 2. To produce this figure, I created separate files for the chapters and analyzed them with cluster analysis; to make the graph more readable, I labeled the chapters 'pre' (preliminary)-'ten,' '11'-'20,' and 'twenty-1'-'thirty-1'. The two main clusters in Fig. 2 do not strongly reflect the structure of the novel: most early chapters are in the left cluster and most middle and later ones in the right cluster, but at least three chapters from each group appear in each. Among the dozens of novels I have analyzed, *Maisie* is unusual in how weakly its narrative structure is reflected in the relationships among the chapters (plot is not James's major interest). Analyses based on the 100, 200, 300, 400, 500, 600, 700, 800, 900, and 990 MFW all fail to suggest any change-point associated with James's adoption of dictation.

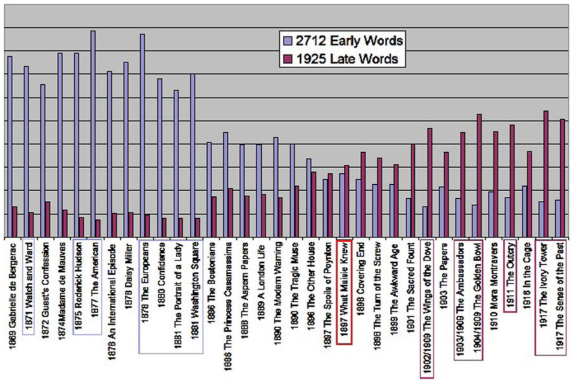


Fig. 1 Early and late words in Henry James's early and late novels and novellas

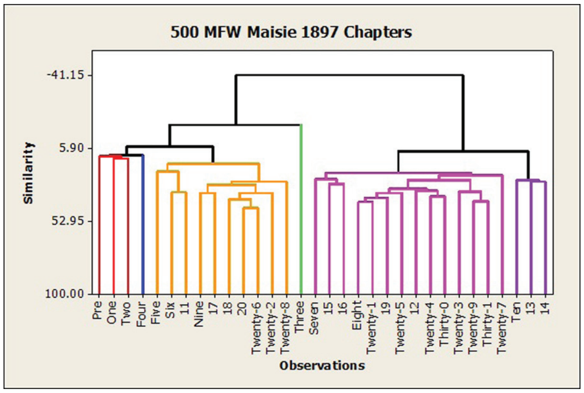


Fig. 2 A Chapter analysis based on the 500 MFW of What Maisie Knew

Consider some other statistics by chapter (see Fig.3). Sentence-length is an obvious choice, given the standard view of James's style, and, predictably, it is closely related to the percentage of narrative. Some variables sharply increase and decrease, but without any obvious change-point. Furthermore, average sentence length trends irregularly lower throughout the novel, as does the percentage of narrative, suggesting, unsurprisingly, that James uses shorter sentences in dialogue, but conflicting with the idea that dictation caused long, rambling sentences. As Fig. 4 shows, average word length and the frequency of long words also trend lower, but there is nothing to suggest a change caused by dictation.

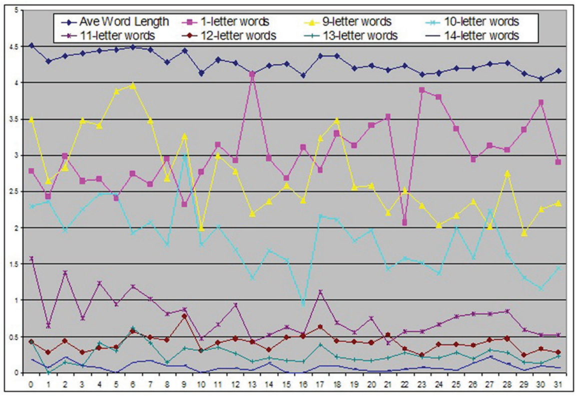


Fig. 3 Vocabulary, sentence-length, % narrative, and word-length in chapters of Maisie

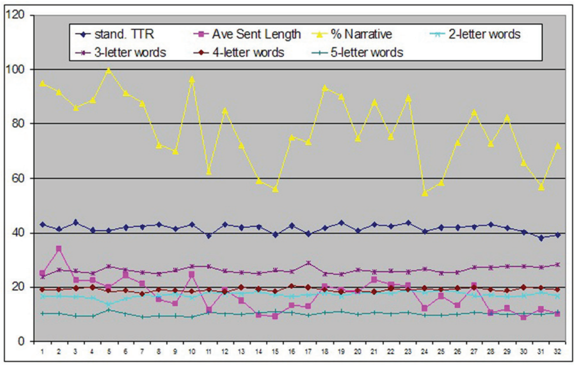


Fig. 4 Ave. word-length, one-letter words, and long words in chapters of Maisie

The next two figures show the distribution of punctuation across the chapters. There are obvious correlations between some punctuation marks and the percentage of narrative, and some surprising correlations (why are there more commas in narrative?). Nothing, however, suggests any shift caused by a change in the mode of composition.

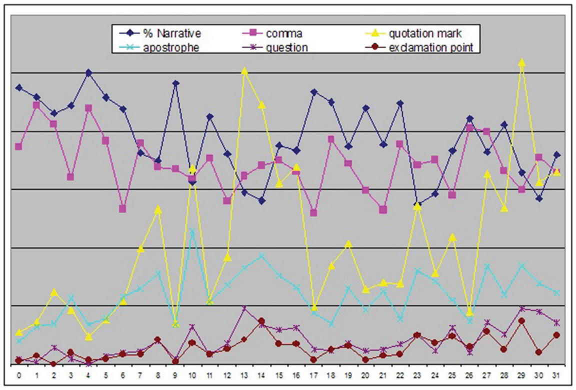


Fig. 5 Some punctuation marks and percent of narrative in chapters of Maisie.

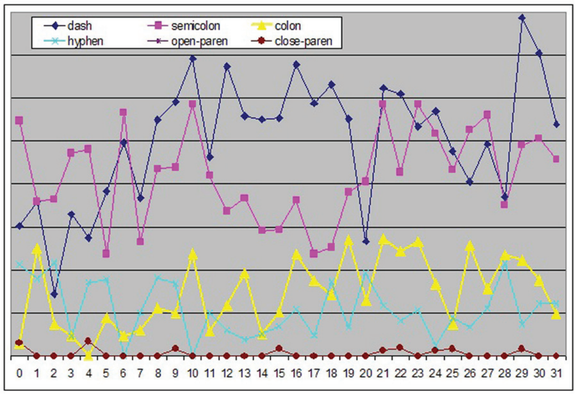


Fig. 6 Other punctuation marks in chapters of Maisie

Finally, Fig. 7 displays vocabulary richness, sentence-length, and the percentage of short words in James's novels over time, and Fig. 8 shows just how regularly the frequency of frequent words characterizes James's chronological development. The patterns shown in these graphs are difficult to reconcile with any significant stylistic change caused by James's adoption of dictation in 1896. (The slightly anomalous position of *The Outcry* in Fig. 8 may be related to the fact that it is a novelization of a play.)

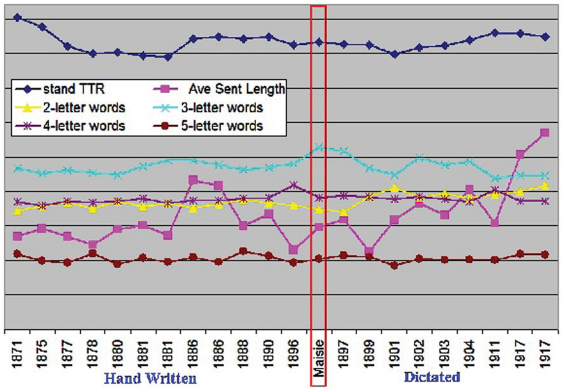


Fig. 7 Vocabulary richness, sentence-length, and word-length in James's novels.

In conclusion, the gradual, unidirectional nature of the changes in James's style does not provide support for any radical alteration caused by dictation. As reasonable as this idea has seemed, and as productive as it has been for speculation about media, machines, and literary production, such speculation may have to be revised or abandoned in the light of new kinds of evidence–evidence only available by computational examination of a corpus of James's work. If further research confirms these preliminary results, computational stylistics may have an important contribution to make to literary studies in this case.

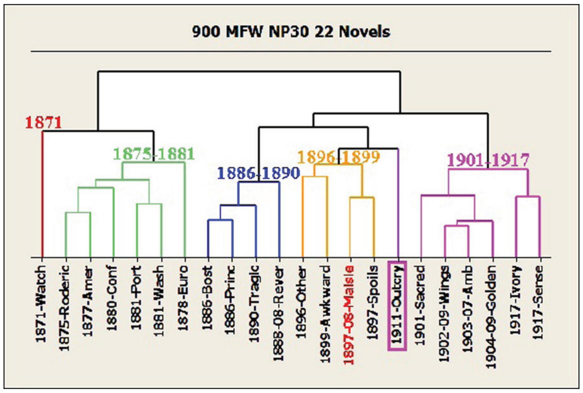


Fig. 8 Chronological development in the style of Henry James

References

Barron's Notebooks. The Turn of the Screw. (1986). Barron's Educational Series, Inc. Online. Available: http://www.pinkmonkey.com/booknotes/barrons/turnscr2.asp. Accessed 11/6/08.

Bosanquet, Theodora. (1924). Henry James at Work. London: Hogarth Press.

Cappello, Mary. (2007). Awkward: A Detour. New York: Bellevue Literary Press.

Wikipedia. (2008). Henry James. Online. Available: http://en.wikipedia.org/wiki/Henry_James.Accessed 11/4/08.

Edel, Leon. (1969). Henry James: The Treacherous Years, 1895-1901. Philadelphia, Lippincott.

Honeycutt, Lee. (2003). Researching the use of voice recognition writing software, Computers and Composition 20(1): 77-95.

Hoover, David L. (2007). Corpus Stylistics, Stylometry, and the Styles of Henry James, Style 41: 174-203.

Bethany [Nowviskie]. (2002). The Ivanhoe Game, 'The Turn of the Screw.' Online. Available: http://speculativecomputing.org/greymatter/ivanhoe/roles/archives/00000019.htm Accessed 11/4/08.

James, Henry. (1897). What Maisie Knew. New York: Stone. Online. Available:

http://ia331316.us.archive.org/0/items/whatmaisieknew00jamerich/whatmaisieknew00jamerich_djvu.txt

McLuhan, Eric, and Frank Zingrone. (1995). Essential McLuhan. Concord, Ont.: House of Anansi.

Newman, Dan. (2000). The Dragon Naturally Speaking Guide, 2nd ed. Berkeley: WavesidePublishing. Online. Available: http://lib.store.yaoo.net/lib/sayican/onlinebook.html. Accessed 11/6/08.

Ozick, Cynthia. (2008). Dictation: A Quartet. New York: Houghton Mifflin.

Seltzer, Mark. (1992). Bodies and Machines. New York: Routledge.

Short, R. W. (1946) The sentence structure of Henry James, American Literature, 18(2): 71-88.

Thurschwell, Pamela. (2001). Literature, Technology and Magical Thinking, 1880-1920. Cambridge: Cambridge University Press.

Co-word Analysis of Research Topics in Digital Humanities

Xiaoguang Wang
Ritsumeikan University
whu_wxg@126.com

Mitsuyuki Inaba
Ritsumeikan University
inabam@sps.ritsumei.ac.jp

Introduction

Over the last two decades, digital humanities has become increasingly popular. Trying to combine computing with traditional disciplines of arts and humanities, digital humanities researchers have come from various fields such as history, philosophy, linguistics, literature, arts, and archaeology. While the interdisciplinary aspect of digital humanities is obvious, its disciplinary structure has not been established yet. In order to identify hot topics and map the disciplinary structure, we made co-word analysis based on a group of research papers' titles.

Related Works

Co-word analysis is a content analysis technique that uses patterns of co-occurrence of pairs of terms, i.e., words or phrases, in a corpus of texts to identify the relationship between these terms, the extent to which these themes are central to the whole area, and the degree to which these themes are internally structured (Qin, 1999). It does not rely on any a priori definition of research themes in science. This enables us to follow actors objectively and detect the structure of science without reducing them to the extremes of either internalism or externalism (Callon et al., 1986).

Method and Data

The first step of co-word analysis is to extract keywords from records in corpus. We chose 516 papers written in English from two journals and four conference proceedings: *Literary and Linguistic Computing* from Year 2005 to 2008, *Digital Humanities Quarterly* from Year 2007 to 2008, and proceedings of *ACH/ALLC 2005, DH 2006, DH 2007*, and *DH 2008*. Since these papers have no author keywords, we manually extracted keywords from their titles and picked out 1231 distinct keywords, which appeared 2040 times in total.

Terms	Freq.	Terms	Freq.	Terms	Freq.
text	44	model	10	database	6
digital	37	translation	10	delta	6
digital humanities	24	collaborative	9	evaluation	6
TEI	20	history	9	framework	6
corpus	19	learning	9	interface	6
archive	16	manuscript	9	language	6
linguistics	15	poetry	9	medieval	6
authorship	14	text mining	9	novel	6
historical	14	gender	8	stylistic	6
digital edition	13	music	8	teaching	6
humanities computing	13	semantic	8	cultural heritage	5
online	13	automatic	7	digital library	5
encoding	12	resource	7	digitization	5
literary	12	case study	7	early modern	5
xml	12	computing	7	mining	5
art	11	electronic	7	reading	5
dictionary	11	French	7	speech	5
English	11	meaning	7	virtual	5
markup	11	metadata	7	visual	5
scholarship	11	variation	7	integration	5
text analysis	11	American	6	knowledge	5
visualization	11	annotation	6	representation	5
web	11	author	6		
attribution	10	culture	6		

Table 1 Top 70 high-frequency keywords

We conducted co-word analysis with two aims. One is to detect the structure of a research field, and the other to detect minor but potentially growing areas. To accomplish these two aims, we selected top 70 keywords (5.68%) whose frequencies were higher than 4 (see Table 1).

These keywords' total frequency was 681 (33.38%).

We calculated the association values between any word pairs with Equivalence Coefficient index (*E*) which can be defined as follows:

$$E_{ij} = \frac{C^2_{ij}}{C_i \times C_j}$$

C_{ij} is the number of documents in which the keyword pair (*i* and *j*) appears. C_i and C_j are the occurrence frequencies of Keywords *i* and *j* in the group of the articles respectively. E_{ij} has a value between 0 and 1. E_{ij} measures the probability of Keyword *i* appearing simultaneously in a document set, indexed by Keyword *j*, and vice versa, given the respective collection frequencies of the two keywords. Therefore, E_{ij} is called "a coefficient of mutual inclusion" by Turner et al. (1988).

Results

Based on the Equivalence Coefficient indexes, we constructed a co-occurrence matrix, and then made a multidimensional analysis.

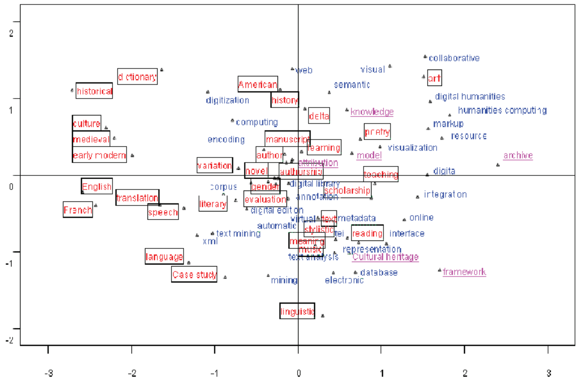


Fig. 1 Multi-dimensional analysis of keywords

Combining the keywords' frequencies with the multidimensional analysis (see Fig. 1), we can see that the keywords in digital humanities can be divided into three categories. The keywords that fall under the first category—the ones without underlines or frames—are directly related to information technologies, such as digital, TEI, XML, web, online, visual, and metadata. The hot topics in this category include visualization, markup, text mining, annotation, and digital library. The keywords that fall under the second category, the ones framed, can be associated with traditional humanities including literary, linguistics, history, culture, language, poetry, novel, and speech. With support of the information technologies mentioned above, many researchers have focused on medieval and early modern literatures, translation between literary works in different languages, authorship, gender, and stylistics. The third category with the keywords being underlined consists of some general words, such as knowledge, model, and framework.

Figure 1 shows these three categories all mix together, with no obvious clusters in them. This means digital humanities is still a new discipline without any subdisciplines formed. What should be particularly noticed is that English and French have been studied more than other languages. This indicates that digital humanities research has made uneven progress, depending upon languages.

A co-word network was drawn with the co-occurrence matrix (see Fig. 2). We calculated nodes' degree centralities in the co-word network with social network analysis method. Degree centrality means the number of co-occurrence with other keywords. As Figure 2 shows, all the keywords can be divided into three levels, according to their degree centralities.

An interesting phenomenon in Table 1 and Figure 2 is that the frequencies of "digital humanities" and "humanities computing" are high, but their degree centralities are low, while most of the others are coessential. A major reason seems that digital humanities research has been furthered in recent years. Six fundamental concepts have been found which may benefit from the dissemination of technologies related to textual digitalization. Still, digital humanities is a new discipline, and researchers have been trying to integrate digital technologies with traditional arts and humanities. Some recent papers' titles include these two keywords and some other low frequency keywords that may indicate new research directions, e.g., geographical information system and interactive games. We could interpret, therefore, these keywords might work as an indicator for the future research topics in digital humanities. While not giving the complete picture of the digital humanities' discipline structure, Figure 2 is complementary to the "intellectual map" painted by McCarty W. and Short H. (2002).

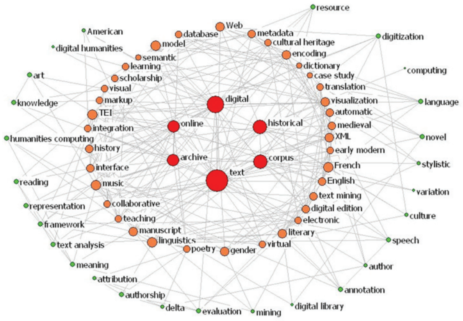


Fig. 2 Co-word network (node size indicates its degree centrality)

Conclusions

In this paper, we analyzed the structure of digital humanities with co-word method. We counted frequencies of keywords which we picked out from the titles of the selected journal papers, recently published. Then we made a multi-dimensional analysis and a network analysis. As a result, six fundamental concepts are found in digital humanities, but there are no clear subdisciplines in it yet.

While a widely used method in library and information science, co-word analysis is still new for digital humanities researchers. In the future, we will improve this analysis further more in terms of methodology, developing an integrated software for its common applications in digital humanities.

References

Callon, M., Law, J., and Rip, A. (1986). How to study the force of science. In Callon, M., Law, J., and Rip, A. (eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world*. London: The Macmillan Press Ltd, pp. 3-15.

McCarty, W. & Short, H. Mapping the field. (2002). http://www.allc.org/content/pubs/map.html

Qin, H. (1999). Knowledge discovery through co-word analysis. *Library Trends*, 48(1):133-159.

Turner, W. A., Chartron, G., Laville, E., and Michelet, B. (1988). Packaging information for peer review: New co-word analysis techniques. In Van Raan, A. F. J. (ed.), *Handbook of quantitative studies of science and technology*. Netherlands: Elsevier Science Publishers, pp. 291-323.

Testing Authorship in the Personal Writing of Joseph Smith Using NSC Classification

Matthew Jockers
Stanford University
mjockers@stanford.edu

Overview

In a co-authored paper published in *Literary and Linguistic Computing* (2008), my co-authors and I employed both delta and Nearest Shrunken Centroid (NSC) classification in an authorship analysis of the *Book of Mormon*. Our results suggested that several men involved in the early formation of the Mormon church were likely contributors to the *Book of Mormon*. For reasons detailed below, we were unable to include Mormon prophet Joseph Smith in our authorship tests. The work presented here attempts to develop a sizeable Smith corpus by using a small set of documents in his own handwriting as a training model for evaluating other documents attributed to Smith but written in the handwriting of one of Smith's 24 different scribes.

For the aforementioned article, we compiled a corpus of source material from five candidate authors who were involved in the early LDS church. We had hoped to include Joseph Smith as a candidate, but in the course of our research determined that there was not enough reliable writing by Smith to constitute an ample sample for testing of his signal in the *Book of Mormon*. As Smith biographer Dean Jessee makes clear in the introduction to *Personal Writings of Joseph Smith*, Smith's speeches, letters, and even journal entries were frequently written by scribes or written in tandem with one or more of his collaborators. In another article that appears in the pages of the "Joseph Smith Papers" online archive (n.d.), Jessee writes, "only a tiny proportion of Joseph Smith's papers were penned by Smith himself." In many of the documents Jessee has collected, we see the handwriting of Smith interwoven with the handwriting of his scribes, sometimes side by side in the exact same letter, journal entry, or document.

Mormon history informs us that Smith frequently used scribes and that he dictated his thoughts to them. Indeed the entire *Book of Mormon* is said to be a verbatim transcript of Smith's dictation. With regard to documenting his visions, thoughts, and experiences, Smith's "philosophy" writes Jessee, "was that 'a prophet cannot be his own scribe.'" That said, on some occasions Smith did put pen to paper, sometimes alone and sometimes in tandem with others. Though Jessee has "attributed" the spirit and content of all of these documents to Smith, the manuscripts show clear physical evidence of other hands at work; thus, the question remains as to whether these scribes were "authoring" or merely "transcribing."

These manuscripts, though not reliable for use as samples of determined authorship in our prior research, do provide fertile ground for another sort of closely related stylistic inquiry and allow an opportunity to investigate the question of whether Smith's various scribes may have contributed more than simple transcription. For this new research I utilize the models of known authorship we developed in our prior work in order to analyze the personal writings attributed to Smith (but written by scribes).

The goal of this work is to assess the role (if any) that the scribes may have had in shaping the linguistic and stylistic construction of these documents. For example, if sections written in the hand of Sidney Rigdon are classified as being similar to the Rigdon signal in our exisiting model, such as result would suggest that the role Rigdon played in the dictation process was perhaps more than mere scribe. Alternatively, if the material not in the hand of Smith is classified together with material that is in his hand, then this would be evidence favoring attribution to Smith and Smith alone.

At the time of this proposal, preliminary results indicate that at least 20 of the 109 documents penned by Smith's scribes are stylistically close to those written in Smith's own hand. Furthermore, early results also suggest no apparent stylistic connection between those scribes for whom we have known writings and the documents that they wrote in the role of scribe to Smith. Together, these findings appear to support the historical Mormon church perspective of common authorship for both the papers in Smith's hand and those in the hands of his many scribes.

Further tests, to be completed before presentation of this research, are necessary to confirm the veracity of these preliminary results. Should further study confirm the preliminary data, then we would have some justification for attributing a sizeable number of these scribe-written texts to Smith. Providing additional authentication of the Smith corpus in this manner would be of great value to future studies of the *Book of Mormon*.

Methodology

I begin by segmenting the personal documents attributed to Smith according to differences in handwriting. For documents not in Smith's handwriting, I label them based on Jessee's identification of the scribe who took the dictation. From the material in Smith's own hand, I expand our current classification model to include a new "Smith" class (the current model includes signals for Oliver Cowdery, Sidney Rigdon, and Parley Pratt who were among Smith's known scribes).

Through cross-validation (and testing with various tuning parameters) I determine the most effective number of features for a new model. In our prior work, NSC had a cross-validation error rate of just 8.8% when using 110 features; the model accurately classified known samples 91.2% of the time. I anticipate similar results for this proposed research.

Using the new model, the corpus of personal writings attributed to Smith, but not in his hand, will be classified and the results ranked based on the probabilistic output that NSC provides. The results will provide further evidence as to the consistency of the linguistic signal across the corpus and provide a foundation for further research.

References

Brodie, F. (1971). *No Man Knows My History; the Life of Joseph Smith, the Mormon Prophet*, New York, Knopf.

Burrows, J. F. (1987). *Computation into Criticism: a Study of Jane Austen's Novels and an Experiment in Method*. Oxford, Clarendon Press.

Burrows, J. F. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17, 267-87.

Burrows, J. F. (2003). Questions of Authorship and Beyond. Computers and the Humanities, 37, 5-32.

Burrows, J. F. (2005). Who Wrote Shamela? Verifying the Authorship of a Parodic Text. *Literary and Linguistic Computing*, 20, 437-450.

Bushman, R. L. (2005). *Joseph Smith: Rough Stone Rolling*, New York, Alfred A. Knopf.

Campbell, A. (1831). An Analysis of *The Book of Mormon* with an Evaluation of Its Internal and External Evidences, and a Refutation of Its Pretenses to Divine Authority. The Millenial Harbinger.

Cowdrey, W. L., Davis, H. A. & Vanick, A. (2005). Who Really Wrote the *Book of Mormon*?, St. Louis, MO, Concordia Pub. House.

Croft, J. D. (1981). *Book of Mormon* 'Wordprints' Reexamined. Sunstone 6, 15-21.

Hilton, J. L. (1988). On Verifying Wordprint Studies: *Book of Mormon* Authorship. Foundation for Ancient Research and Mormon Studies, Provo, Utah 1988. Holley, V. (1989). Book of Mormon Authorship: A Closer Look, UT, Self-published.

Holmes, D. I. (1991a). A Multivariate Technique for Authorship Attribution and Its Application to the Analysis of Mormon Scripture and Related Texts. *History and Computing*, 3, 12-22.

Holmes, D. I. (1991b). Vocabulary Richness and the Prophetic Voice. *Literary and Linguistic Computing*, 6, 259-268.

Holmes, D. I. (1992). A Stylometric Analysis of Mormon Scripture and Related Texts. Journal of the Royal Statistical Society, Series A (Statistics in Society), 155, 91-120.

Hoover, D. L. (2002). Frequent Word Sequences and Statistical Stylistics. Literary and Linguistic Computing, 17, 157-180. Hoover, D. L. (2004a). Delta Prime? *Literary and Linguistic Computing*, 19, 477-493.

Hoover, D. L. (2004b). Testing Burrow's Delta. *Literary and Linguistic Computing*, 19, 453-475.

Jessee, Dean C. (ND). "Joseph Smith and His Papers: An Editorial View." Joseph Smith Papers Project, http://josephsmithpapers.org/Essays/Jessee.pdf (accessed 2/25/2009).

Jockers, Matthew L., Daniela M. Witten, Craig S. Criddle. "Reassessing Authorship in the Book of Mormon Using Delta and Nearest Shrunken

Centroid Classification." *Literary and Linguistic Computing*, 2008; doi: 10.1093/llc/fqn040

Larsen, W. A., Rencher, A. C., Layton, T (1980). Who Wrote the *Book of Mormon*? An Analysis of Word-prints. *BYU Studies*, 20, 225-51.

Reynolds, N. B. (2005). The Case for Sidney Rigdon as Author of the Lectures on Faith. *Mormon History*, 32, 1-41.

Riley, I. W. (1902). The Founder of Mormonism: A Psychological Study of Joseph Smith, Jr., New York, Dodd, Mead, and Co.

Smith, J. & Jessee, D. C. (2002). *The Personal Writings of Joseph Smith*, Salt Lake City, Brigham Young University Press, Deseret Books.

Tibshirani, R., Trevor Hastie, Balasubramanian Narasimhan and Gilbert Chu (2003). Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. *Statistical Science*, 18, 104-117.

Tibshirani, R., Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99, 6567-6572.

Vogel, D. (2004). *Joseph Smith: The Making of a Prophet*, Salt Lake City, Signature Books.

The Social Text as Digital Gamespace; or, what I learned from playing *Spore*

Steven E. Jones
Loyola College
sjones1@luc.edu

The meaning of video games is in their playing, in performance, which involves searching for signs of intelligence—human and machine-mediated—with which to collaborate or compete in making meanings. This happens in textual interpretation of all kinds, but the process is foregrounded and dramatized in video games in ways that make them useful models for textual scholarship, editing, and interpretation, considered in terms of scholarly content management systems or socially networked knowledge sites. For example: what I learned from playing Will Wright's sim-everything game, *Spore*, is the importance of building asynchronous "pollination" systems that encourage complex improvisation by way of feedback loops, afford the search for and collaboration with others (via their creations), the continual re-editing of shared materials, and preserve, track, and allow for the analysis of multiple individual edits.

Interactive Exploration of Versions across Multiple Documents

Chang-Han Jong
University of Maryland
chjong@umd.edu

Prahalad Rajkumar
University of Maryland
prahalad@cs.umd.edu

Behjat Siddiquie
University of Maryland
behjat@cs.umd.edu

Tanya Clement
University of Maryland
tclement@umd.edu

Catherine Plaisant
University of Maryland
plaisant@cs.umd.edu

Ben Shneiderman
University of Maryland
ben@cs.umd.edu

Introduction

The need to compare two or more documents arises in a variety of situations. Some instances include detection of plagiarism in academic settings and comparing versions of computer programs. Extensive research has been performed on comparing documents based on their content (Si et al., 1997; Brin et al., 1995) and there also exist several tools such as *windiff* to visually compare a pair of documents. However, little work has been done on providing an effective visual interface to facilitate the comparison of more than two documents simultaneously. *The Versioning Machine* (Schreibman et al., 2003) is a web-based interface that provides the facility to view multiple versions of a document, along with the changes across versions. Motivated by the Versioning Machine (VM), we build a tool *MultiVersioner* that facilitates viewing multiple versions of multiple documents at once, and provides the user with a rich set of information regarding their comparison. The primary user during the development of MultiVersioner was Tanya Clement, a doctoral candidate in English at the University of Maryland, who researches the works of experimental poets.

Related Work

ScentHighlights (Chi et al., 2005) has demonstrated the effectiveness of using color-coded highlighting to display the similarities and differences across documents. There exist literature (Brin et al., 1995) and tools like *CHECK* (Si et al., 1997) and *MOSS* (MOSS) on plagiarism and source code comparison, which are relevant to our work. *FeatureLens* (Don et al., 2007) facilitates pattern finding in text collections by providing visualizations of the results of text mining algorithms.

In Tanya Clement's research, she compares not only versions of a single poem, but also multiple versions across several poems. VM can display the versions of just one document at a time. To open another document, all versions of the current document have to be closed first. VM also does not provide any search capabilities.

Description of the Interface

Background

The two-fold goal of MultiVersioner is to provide an effective overview of the content and size of all documents, as well as to provide a detailed display, along with a variety of search capabilities, in accordance with the *Info-Viz Mantra Overview first, zoom and filter, details on demand* (Shneiderman 1996).

MultiVersioner is implemented in Java 6.0 using the Swing GUI toolkit. It uses the same input format file as VM, an XML file, containing information about the various additions and deletions made across all versions of a document. MultiVersioner contains a built-in parser to parse these XML files. Loading an XML file opens all the versions of a poem in separate *version panels* and multiple such documents can be loaded simultaneously. Version panels are displayed in the central part of the interface with a tool panel located on the right. The name of the version appears on top of the respective version panel. The names of all the versions of a particular document are displayed in the same color in order to group them together.

Overview

In the overview, words are denoted by equal sized boxes. Hovering over a box pops up a tooltip containing the entire sentence, with the current word being displayed in bold. In the tooltip, words added in the current version are shown in blue, and words deleted are shown in red. Clicking on a box brings up a *detail window* (Figure 1) containing the entire sentence. The purpose of the detail window is to display a sentence of interest on the screen, analogous to a post-it note. The detail windows can be made either opaque or transparent, and can either be moved around, or aligned together. A line is drawn between a detail window and its corresponding location in the version panel, to keep track of its origin. A detail window could be closed either by right clicking it and choosing close, or by simply dragging it out of the screen.

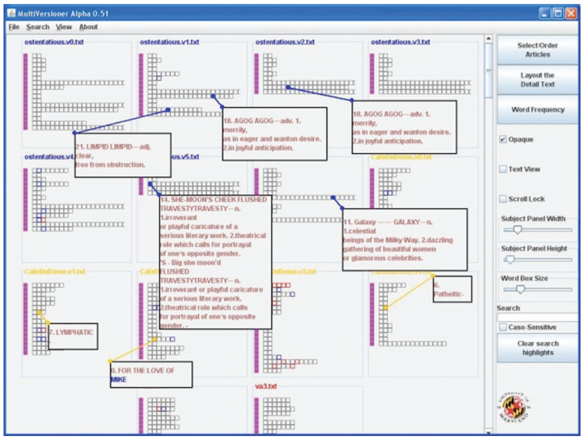

MultiVersioner Alpha 0.51
File Search View About
ostentatious.v0.txt
ostentatious.v1.txt
ostentatious.v2.txt
ostentatious.v3.txt
Select/Order Articles
Layout the Detail Text
Word Frequency
Opaque
Text View
Scroll Lock
Subject Panel Width
Subject Panel Height
Word Box Size
Search
Case-Sensitive
Clear search highlights
21. LIMPID LIMPID—adj. clear, free from obstruction.
18. AGOG AGOG—adv. 1. merrily, as in eager and wanton desire. 2.in joyful anticipation.
14. SHE-MOON'S CHEEK FLUSHED TRAVESTYTRAVESTY—n. 1.irreverant or playful caricature of a serious literary work. 2.theatrical role which calls for portrayal of one's opposite gender. - Big she moon'd FLUSHED
11. Galaxy —— GALAXY—n. 1.celestial beings of the Milky Way. 2.dazzling gathering of beautiful women or glamorous celebrities.
7. LYMPHATIC
9. FOR THE LOVE OF MIKE
Pathetic
va3.txt

Fig. 1 Overview of multiple versions of two poems. Sentences of interest are shown using detail windows, which are linked to their respective version panels by lines.

Text View

Word boxes are used primarily to obtain an overview of all the documents. To explore the versions in detail, a representation displaying the actual sentences, instead of word boxes, is preferred (Figure 2).

Search

The basic search feature is the word search (Figure 2). A search bar is provided where the user can type in a word or a phrase to be searched across all documents. A search can be made case-sensitive if desired. Alternatively, a word can be searched by right-clicking an instance of it. Inspired by ScentHighlights (Chi 2005 et al., 2005), Search results are color-coded. The instances of a searched word in all documents are highlighted using the same color. A search history as well as the facility to clear search results is available.

A line search feature is available as well. The similarity of a pair of lines is computed by taking into account the number and relative positions of the words common to them. Right-clicking the anchor-box present at the beginning of a line triggers a line search and lines similar to the specified line across all documents are highlighted.

Word Frequency Table

MultiVersioner computes a frequency table containing the number of occurrences of each unique word in all documents and their versions. When comparing different versions of a document or comparing different documents that are related, researchers in literature need to identify unique and common words and sentences. It has been shown that an approach as simple as a frequency table listing is powerful in providing insight by letting users know which words are common across documents and which ones are unique to a single document (Filippova, 2007).

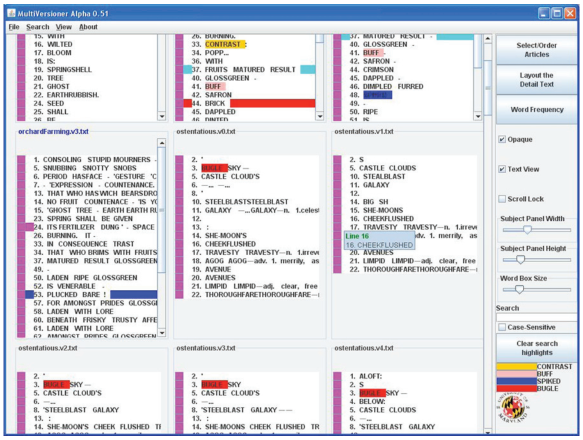


Fig. 2 A text view of versions of two poems. Four different searches for the words "contrast", "buff", "spiked" and "bugle" are performed across all versions of both poems. Each instance of searched word is highlighted using the same color in all documents.

Other features

There are sliders available to control the version panel height, width and the sizes of the word boxes. MultiVersioner also has a scroll lock, used when the documents are long, to synchronize the scrolling of the documents with each other.

User Feedback and Future Work

The first three authors of the abstract were the developers (students in an Information Visualization class). The last three authors used the prototypes several times during the development and provided suggestions. Examples are given below.

Users reported that they could not easily distinguish versions of a single document from versions of other documents. We achieved this by using the same color for the titles of all the version panels associated with a the same document. The ability to search for words across the documents was greatly appreciated and the color-coded highlighting of the results was found helpful, but overall they commented that softer colors would be more pleasing. Users liked the synchronized scrolling between windows. Still it was difficult to associate the detail window to its originating location, so some visual linking was suggested (e.g. drawing a line or matching color highlighting). Regarding the choice between Text View and Overview, users stressed that it is preferable to see the actual words rather than seeing the abstract word boxes. Still, they acknowledged the benefit of the overview when dealing with a large number of documents and versions. The overview was more helpful for a high-level view and orientation, the text view was more useful for analysis.

In summary we believe that our prototype illustrates that it is possible to facilitate visual comparison. We built on the Versioning Machine by allowing users to compare multiple documents, each of which consisting of multiple versions. We also provide the ability to search for entities such as words and lines across the documents and versions and analyze their frequency patterns. MultiVersioner was designed to compare small poems, and future work needs to address the problem of longer documents. Utilizing the entire screen space, by dynamically resizing all open documents to fit the screen, might be a promising direction. Further information about this project including a more detailed report, slides, a demo video and software can be found at https://wiki.cs.umd.edu/cmsc734_08/index.php?title=MultiVersioner.

References

Brin, S., Davis J., and Garcia-Molina H. (1995). *Copy detection mechanisms for digital documents*, Proc. of the 1995 ACM SIGMOD international conference on Management of data, 398-409.

Chi, E. H., Hong, L., Gumbrecht, M., and Card S. K. (2005). *ScentHighlights: highlighting conceptually-related sentences during reading*, Proc. of the 10th International Conference on Intelligent User Interfaces, 272-274.

Don A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., and Plaisant, C. (2007). *Discovering interesting usage patterns in text collections: integrating text mining with visualization*, Proc. of the sixteenth ACM Conference on Information and Knowledge Management, 213-222.

Filippova, D. (2007). *BasketLens: interface for document visualization and exploration*, http://www.cs.umd.edu/hcil/textvis/basketlens/.

MOSS. http://theory.stanford.edu/~aiken/moss, retrieved 04-10-2008

Schreibman, S., Kumar, A., and McDonald, J. (2003). “The Versioning Machine.” *Literary and Linguistic Computing*, 18(1), 101-107 (http://www.v-machine.org)

Shneiderman, B. (1996). “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization.” *IEEE Conference on Visual Languages*, 336-343.

Si, A., Leong, H. V., Lau, R. W. H. (1997). CHECK: a document plagiarism detection system, *Proc. of the 1997 ACM symposium on Applied computing*, 70-77.

Audio-visual Rhetoric and its Methods of Visualization

Introducing visual notation systems for film analysis

Gesche Joost
Deutsche Telekom AG Laboratories
gesche.joost@telekom.de

Sandra Buchmüller
Deutsche Telekom AG Laboratories
Sandra.Buchmueller@telekom.de

Tom Bieling
Deutsche Telekom AG Laboratories
Tom.Bieling@telekom.de

Abstract

In this paper, we use information design of rhetorical theory in order to investigate the structure of audio-visual media. From a design research perspective, we are especially interested in the power of visualization. In this respect, we compare different methods of film analysis referring to their representation mode and evaluate them referring to their ability to visualize the cinematic structure. Within the realm of film analysis, we ask :

- How can the use of information design enhance the traditional analysis of film?
- What is the influence of the selected semiotic code (text, image, or animation) on the interpretation of the medium and its structure?

To answer these questions, we introduce examples of visual notation systems that were applied to analyze film sequences within a project. We discuss their method and value for the purpose of analysis and reflect on the role of the different semiotic codes that are applied. Referring to the traditional method of analysis in the form of written film protocols, we hypothesize that this method is attended by a high loss of structural information when transferring the audio-visual, dynamic code of film into written text. While film protocols are more appropriate for content, respectively plot analyses, we state that static and dynamic visual notations are suitable forms for making structural and rhetoric phenomena of film explicit. Visual notations might cause the least loss of information in the analytical process, and might of high value for academic film analysis.

Background

Audio-visual rhetoric is a knowledge domain for designers in theory and practice that was introduced by Gui Bonsiepe in the 1960ies (Bonsiepe 1961). Its theory and communication system is valid for all communicative actions aiming for *persuasion*. Richard Buchanan described the whole body of design practice as rhetorical argumentation (Buchanan 1989) and opened the way for a rhetorical design theory on a broad basis. In the following, visual rhetoric was established within the design education mostly in the US, teaching the analysis of information design on the basis of rhetorical patterns (Ehses 1984, 1986, 1988), (Kostelnick 1989, 1998, 2003), (Poggenpohl 1998). These concepts were transferred to dynamic, audio-visual media by Bonsiepe in the late 1990ies, and he introduced the term “Audio-visual Rhetoric” (Bonsiepe 2008). Today, this theory combines the ancient communication theory and its huge body of knowledge with New Rhetoric (Joost 2008, 2008a) and applies this knowledge to the design domain and its visual approaches.

The Notation System for Film Analysis

Exploring audio-visual patterns of film is an analytical endeavor that is still in search for a useful method. In film theory and practice, there are different models to describe and analyze filmic structures. For film scholars, the most established method is using a written film protocol to elaborate on formal aspects as well as on narrative structures. This method was questioned by many scholars because of the lack of potential to reflect on the specific dynamic and audio-visual quality of film (Lehner 1987). Audio-visual rhetoric provides an approach to overcome this problem by introducing a notation system for film. In academic context, there are some examples of visual film protocols (Hahne 1992), (Ramsbott / Sauter 1988). Nevertheless, none of the visual systems has been established for film analysis in a broader context and has not yet reflected the progressing nature of film. A general issue is that there is no interdisciplinary collaboration between study of film, film production, rhetoric, and design research to come up with an applicable system. Bringing together all these competences, one could design a comprehensive system that could be used in various contexts – academic as well as applied. In this context, we discuss the impact of the semiotic code on the film analysis (Table 1).

Comparison between different notation systems for analysis

When applying rhetorical knowledge to the analysis of film, we take the following steps. We raise the hypothesis that a visual analysis of film conveys a *visual knowl-*

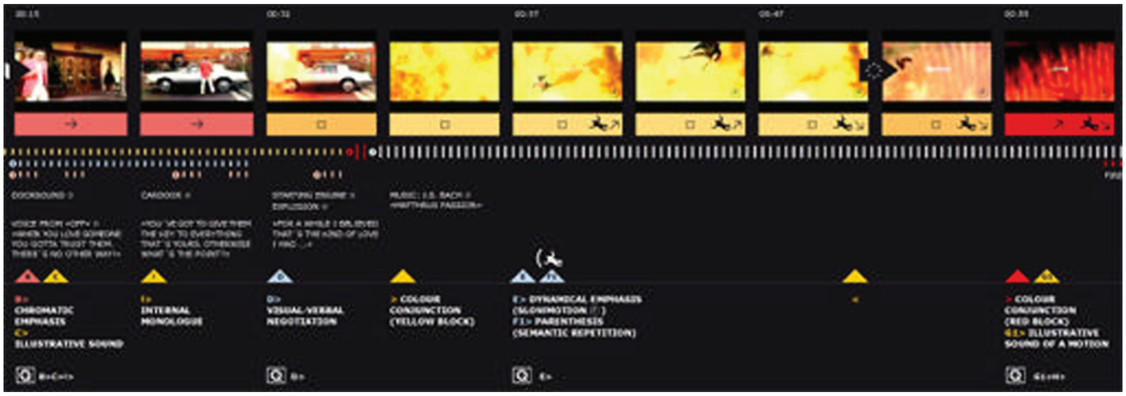


Fig. 1: Static visual Film Protocol

edge about the rhetorical structure of media. This form of knowledge can hardly be expressed verbally. Ernst Gombrich claims what the diagram presents in front of our eyes can hardly be expressed by words – as a succession of statements (Gombrich 1984). An example for this hypothesis is the visualization of the spatial information on a map. This complex information that is accessible parallel on one sight can hardly be expressed verbally in a cognitively efficient process. Based on this hypothesis, we introduce different notation systems to visualize recurring patterns in film: a static visual film protocol (Fig. 1), an animated film protocol (Fig. 2), and a written text protocol.

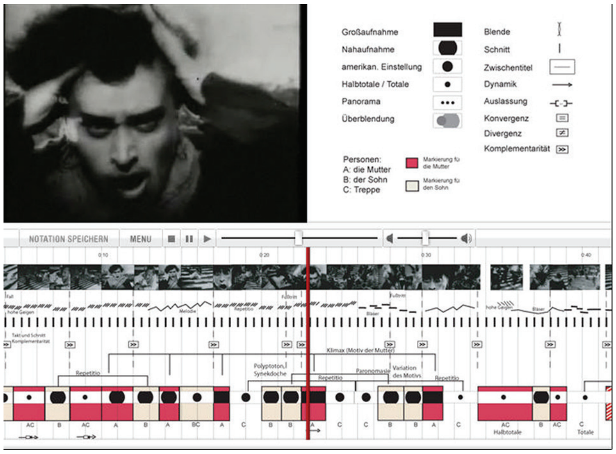


Fig. 2: Animated Film Protocol http://www.geschejoost.org/AVRhetorik/

In the standard film protocol, the translation of audio-visual signs into a written text means to change the semiotic code in a radical way – from image and sound to text. This process involves a loss of information, particularly of the audio-visual and dynamic quality of the sign system. The idea of using a visual protocol as tool for film analysis is based on the hypothesis that a visual diagram can be processed cognitively much more efficient than language (Bonsiepe 2008). This is especially true for the visual aspects of film and not for re-telling its storyline. It is not new to say that the tools and methods that are used for an investigation clearly influence the research results. This is also true for the method we suggest in this paper. To set up a visual diagram of film focuses much more on visual and structural aspects than on the storyline. The aim is to reduce a loss of information that occurs when the audio -visual texture of film is transferred to written film protocol. Therefore, the audio-visual signs are translated into a visual structure. This method has additional advantages: the graphic displays information on one sight so that the recipient can process the data in parallel. Written text can communicate information only in a successive way – one word after the other. In the notation protocol, information about the whole clip in each of the channels can be visualized at the same time, allowing a parallel interpretation of data and of relationships among the various audio -visual elements. Here, the pattern structure can be easily identified on the basis of a graphical representation. For example, repetitions or climax patterns can be singled out quickly on the basis of their visual form. With this approach, large amounts of data from audiovisual media can be efficiently processed for analytical purposes. In the next step, we compared the different systems according to their semiotic code, their mode of information and the kind of perception they require.

	Static visual Film Protocol	Animated-visual Film Protol	Written Film Protocol
Sign system (Zeichensystem)	image, language	image, sound, language	language
Semiotic code	icon, index, symbol (Peirce)	icon, index, symbol (Peirce)	symbol (Peirce)
Change of media	film to image	film to filmA	film to language
Reception	synchronic	synchronic and successive	successive

Mode of information	spatial	spatial, temporal and acoustic	temporal
Link to rhetoric	evidentia (evidence)	analogy	argumentation

Table 1: Comparison of different Methods of Film Analysis

Discussion

One finding is that a verbal analysis, on the one hand, is much more focused on content and storyline of a film and translates the audio-visual medium into a verbal narration. It uses the symbol as semiotic code and requires a successive perception. Its mode of information is temporal. The visual analysis, on the other hand, is more focused on the visual structure of film and its recurring patterns. It uses icon, index, and symbol as semiotic code and requires a synchronic perception. Therefore, it displays spatial information. An animated protocol adds dynamic information about the development of film and includes the analyzed film itself into the animation. It also refers to icon, index and symbol in its semiotic code and requires synchronic as well as successive perception. Its mode of information is rich and includes spatial, temporal and acoustic data. Through visual approaches such as static visual protocols as well as animated ones, a different kind of knowledge is gained compared to a mere verbal analysis: a specific kind of *visual* knowledge. This study argues in favor of a semiotic autonomy of visual signs.

Conclusion

The visual film protocol can serve developers of audio -visual media in various ways. First of all, it is a helpful tool to analyze and interpret film and understand its composition. Moreover, the static visual film protocols make different films structurally comparable. Using the notation system one gets a visual protocol comparable to music notations that can be used for reproduction purposes. This leads to the third point: the notation system as a software could be used as tool for film design and planning in addition to the technique of storyboarding, which is still a standard tool for the production process. With this visual aid, film makers can compose their texture beyond sketches of the scene and visual description. Based on these insights, we suggested the notation system as a new tool for film analysis for designers as well as for film scholars.

References

Bonsiepe, G. (1961). Persuasive Communication: Towards a Visual Rhetoric. In Crosby, T. (Eds.) Uppercase Nr. 5, London, UK, 19–34.

Bonsiepe, G. (2008). Audiovisualistische Rhetorik in zeitbasierten Medien: Über die kognitive Relevanz diagrammatischer Visualisierungen. In: Joost, G., Scheuermann, A. Design als Rhetorik. Birkhäuser Verlag Basel. 217-232.

Buchanan, R. (1985): Declaration by Design: Rhetoric, Argument, and Demonstration. In Design Issues, Volume II, Number 1. 4-23; Reprint in: Margolin, V. (1989) (Ed.): Design Discourse. History, Theory, Criticism, Chicago. 91-109.

Gombrich, E. (1984): Representation and Misrepresentation (Reply to Murray Kreiger) In Critical Inquiry, Volume 11, 2, Chicago. 195-201.

Hahne, D. (1992). Komposition und Film: Projekt nach Motiven aus Camus „Der Abtrünnige“ für Chor, Orchester und Spielfilm. (Volkwang-Texte III. Bd. 5. Hrsg. v. Josef Fellsches).

Lehner, C. (1987). Einige zentrale Probleme der neueren Filmsemiotik. In Bauer, L., Ledig, E., Schaudig, M. (Eds.) Diskurs film: Strategien der Filmanalyse. Bd. 1 , München, Germany, pp. 59–72.

Joost, G. (2008): Die rhetorische Pattern-Language des Films. In Joost, G., Scheuermann, A. (Eds.) Design als Rhetorik, Birkhäuser Verlag Basel, 233 – 249.

Joost, G. (2008a) Film-Sprache. Die audio-visuelle Rhetorik des Films. Bielefeld. See also: http://www.geschejoost.org/AVRhetorik/

Ramsbott, W., Sauter, J. (1988). Visualisierung von Filmstrukturen mit rechnergestützen Mitteln. In: Helmut Korte, Werner Faulstich (Hrsg.): Filmanalyse interdisziplinär. (Zeitschrift für Literaturwissenschaft und Linguistik. Hrsg. v. Helmut Kreuzer. Beiheft 15). Göttingen. pp. 156–165.

Conjecture Generation in the Digital Humanities

Patrick Juola
Duquesne University
juola@mathcs.duq.edu

Ashley Bernola
Duquesne University
abernola@gmail.com

Digital scholarship has been very helpful in the development of humanities research (Juola, 2008a), primarily by automating processes such as communication, text processing, and search and permitting scholars to concentrate on analysis and explanation. However, when computers attempt to generate meaning, to perform analysis themselves, the results are usually less than satisfactory and don't actually explain much.

An example of analytic failure can be seen in the reception of "nontraditional" (i.e. statistical, computer-aided) authorship attribution. It is now unquestionable that computers can infer authorship attributes with high accuracy (see Juola, 2008b), but the accurate inference processes tend not to inform us about the actual authors (Craig, 1999). Argamon (2006) has provided a theoretical analysis of one particular method, but in the unfamiliar and ``inhuman'' language of statistics, which again sheds little light on authorial language and authorial thought. By contrast, studies of gender differences in language (e.g., Coates, 2004) offer not only lists of differences, but explanations in terms of the social environment. In fact, the interesting part of scholarship is not in mere observation, but in the refinement and explanation.

This suggests a relatively novel model for computer/human interaction in scholarship, one in which the computer is used to identify patterns that are passed to human experts for validation and explanation. While not widely used, this model has been successfully applied in mathematics by the *Graffiti* program (Fajtlowicz, 1986). This program generates conjectures (randomly) of the form $X < Y$ (or similar forms such as $X < Y + Z$) where X, Y, and Z are "graph invariants," simple numeric properties of graphs such as average number of edges per node, number of colors necessary to color the graph like a map, average distance between nodes, number of different paths between nodes, and so forth. Graffiti then compares this conjecture against a library of graphs to see if it is true for all the graphs in the library. Of course, "true for all the graphs in the library" is not proof of universal truth, but it provides evidence in support. If the conjecture is thus supported, Graffiti publishes the conjecture, and professional mathematicians are invited to prove (or disprove) it. Since inception, Graffiti has developed and published more than 1000 conjectures and inspired more than 100 papers.

This model can easily be applied to the humanities. Application in the field of text analysis is straightforward; we need analogues to "graph invariants." Such "text invariants" might include the frequency of specific words, phrases, or structures in particular text types. A simple conjecture might be that "color adjectives" are more common than "size adjectives," or that "verbs of motion" are more common in male-written novels than in female-written poetry. Of course, text invariants are not limited to token frequency analysis; any "property" that can be assessed via computer analysis would be a possibility. If one can find a way to determine a text's eroticism, degree of animacy, personification, etc., any of those would be potential features.

In addition to text invariants, we need a framework for conjectures (in analogy to the $X < Y + Z$ framework described above). Simple comparisons are likely to find uninteresting conjectures (prepositions are more common than proper nouns). A more interesting conjecture (in the author's opinion) would be something like "color adjectives are more common than size adjectives in women's writing, but the reverse holds in men's writing." Such a finding would clearly show a relationship, as yet unexplained, between adjective choice and gender. If this were true—why? It obviously says something about gender and culture, but what? Here is where a traditional humanist could take advantage of the ability of a computer to read and analyze a huge amount of data very quickly. In general, conjectures of the form "$X > Y$ in texts of category A, but $Y > X$ in texts of category B" (where A and B are non-overlapping categories, ideally pulled from the document's metadata) are likely to be of interest to category A/B specialists, especially if X and Y are themselves interesting natural properties. Another framework would be that "X is more common in A-texts than B-texts," and it is this framework that we use in our prototype.

To illustrate this, we have built a simple version of this conjecture generator ("conjecturator") using standard Java technology, much of it drawn from the JGAAP project (Juola, 2007). The Moby Thesaurus II lists more than 30,000 different synonym sets: for example, the word group "raft" (as in "a raft of money") includes words/phrases such as "barge," "boat," "pile," "pot," and "quite a little." The word group "take back" includes terms like "abjure," apologize," "renege," "disown," and "nullify." We have also collected eight separate (English) translations of the Bible ranging from the Authorized (King James) Version to Revised Standard Version and the Bible in Basic English. Our program selects one synonym set and two Bible versions (at random), then counts every appearance of each word token listed in the synonym set. Our conjectures are therefore of the form:

"Words in `<this category>` appear at least 50% more frequently in `<this>` Bible translation than in `<that>` one."

Our prototype strips punctuation and case differences, but does not perform morphological analysis or even word sense disambiguation. Despite this limitation, simple word-counting reveals that the word group "take back" appears approximately twice as often in the RSV as it does in Young's Literal Translation. We have similarly found that the word group "rhythmical" occurs substantially more often in the KJV than in Darby's Translation. We have at this writing no explanation, but offer them (along with many other findings) to interested Biblical scholars as a potentially unexplored facet of the differences among different translations. (A list of conjectures will be available both electronically and at the conference—about 40% of conjectures appear to be valid, a percentage we find surprising.)

Testing these conjectures has been relatively easy (if time-consuming); We simply attach the program to a large database (in this case, of Bibles) and allow it to sample from the database until it has either confirmed or rejected the hypothesis to its satisfaction. (For example, there does not appear to be a significant difference in the word group "unauthorized" between the RSV and the American Standard Version.) We can extend such a program even to help solve the "how do you read a million books" problem, since the program could not only do the bulk of the initial reading to see if the hypothesis is true in the first place, but would automatically generate a reading list for scholars interested in following up on the conjecture. (At a second per book, a computer could analyze all million texts and deliver a list of how each work fared vs. the conjecture in less than two weeks. By contrast, a human closely reading one book per day would require 3000 years to read a million books.) Even our prototype system can examine many more categories and hypotheses than even the most avid and interested human reader—in its first 24 hours alone, it found more than 200 possibly interesting differences between Bible versions.

As a further extension, we have extended the Conjecturator to include multiple documents and a more robust form of statistical analysis. Using a collection of more than 100 Victorian novels (courtesy of David Hoover, NYU), we now observe mean word usage within a group such as bildungsromanen or gothic novels, compute variance and t-statistics and accept a conjecture if the computed p-value is sufficiently low (in either direction). Results of this further experiment will also be presented.

Although our prototype is limited to text analysis, the possibility of automatic conjecture generation may extend further. A large and rich database of GIS and/or census information may be able to support, for example, conjectures of the form "`<Object A>` is more common in `<Environment X>` than `<Environment Y>`." An example of such a conjecture would be a relationship previously unimagined between the number of veterinarians and Methodist churches in coastal counties.

What are the benefits of such a program? This conjecture generator can deliver a set of (partially) validated observations about easily observable, superficial properties of the texts in the library (or points in the database more generally defined). By construction, all published conjectures are more or less guaranteed to describe something true, at least about the library. These partial truths, to humanists interested in the study of the library, may represent insights that they have not considered and a the particular hypothesis under study. Indeed, the scholars may lack the time to familiarize themselves with every volume in the library, and may not even be "digital" enough to understand the computer analysis, but who may be interested enough to see out the new material that they now know is there.

At the simplest possible level, 1000 validated conjectures are 1000 topics for student projects, research papers, or Ph.D. theses, a partial solution to the "I need to do a term paper but don't know what I want to do it on" question that plagues all supervisors.

More generally, however, this program would also allow humanists to concentrate their efforts on what is generally the most interesting and rewarding part of humanities research; the search for an explanatory theory of human behavior. By giving scholars a list of statements that are probably true, they can concentrate their efforts on producing statements and theories that are genuinely meaningful.

Cross-linguistic Transference of Authorship Attribution, or Why English-Only Prototypes Are Acceptable

Patrick Juola
Duquesne University
juola@mathcs.duq.edu

Authorship Attribution (Juola, 2008) can be defined as the inference of the author or her characteristics by examining documents produced by that person. It is of course fundamental to the humanities; the more we know about a person's writings, the more we know about the person and vice versa. It is also a very difficult task. Recent advances in corpus linguistics have shown that it is possible to do this task automatically by computational statistics.

A key question, however, in any statistical study (not just of text statistics) is whether the data or methods will transfer from one domain to another. Statistical analyses hinge on assumptions which may or may not be met by different languages, and of course, data which is representative of one domain is highly unrepresentative of any other, by definition. A method that performs well in one area may fail miserably in another. As an example, a part-of-speech tagger with 96% accuracy on newswire data may achieve only 50% on chat logs. (Craig Martell, p.c., 2008)

In light of this finding, cross-problem transferrence can be a major problem for statistical authorship attribution. Should we expect an authorship attribution system that performs well on, say, English, to also perform well on Dutch, Serbian, or Chinese? Alternatively, is it reasonable for a scholar of, say, Polish poetry to have confidence in methods that have been tested to perform well, but only on English documents? This, of course, is a major problem, especially with problems of "forensic" interest where the accuracy rate is one of the major considerations regarding the evaluation or even admissibility of evidence.

The JGAAP software framework (Juola et al., 2006) in conjunction with the Ad-hoc Authorship Attribution Competition corpus (Juola, 2004) provide us with some preliminary results. JGAAP (described elsewhere in this volume) is a modular, Java-based program capable of performing thousands of different types of authorship attribution methods on a well-defined corpus. The AAAC comprises 13 authorship attribution problems in a variety of languages and genres. This setup makes large-scale performance comparisons among categories practical.

For example, 8 of the 13 AAAC problems involved English text (in some form or another), but 5 involved other languages such as French, Dutch, Latin, or Serbian/Slavonic. If authorship attribution did not transfer well, we would expect to see little correlation between the average performance of a method on English texts and its performance in other languages, as high-performing methods would not necessarily remain high-performing in other environments. Conversely, if we see a high degree of correlation across languages, this argues for a high degree of transfer.

As part of some other large-scale technical comparisons (this volume), we have gathered 281 separate analyses of the AAAC data using a variety of preprocessors, event set models, and analytic methods, ranging from simple lexical statistics or nearest neighbor histogram measures to complex machine learning models such as support vector machines on word or character n-grams of various sizes. In this database, the correlation between a method's average performance in English and non-English was 0.6680, a highly significant ($p < 0.0001$) result. More tellingly, the coefficient of determination (r^2) was 0.4462, meaning that approximately 45% of the variation in performance of an algorithm across non-English data could be explained simply by a measure of its performance on English-only data; variations in genre, register, or even variations across the broad category of languages-that-aren't-English have only a little more effect in total. (We should add that work is ongoing and the 281 analyses will undoubtedly be expanded in the next several months; these numbers are therefore only preliminary and updated results will be presented.)

Similarly, we can divide the AAAC problems into "large" problems (those with training samples of more than 100,000 characters each) and "small" ones. (Of the 8 English problems, 4 were large; of the five non-English problems, 3 were large.) Across the same 281 analyses, the correlation between performance on large problems and small problems was again highly significant (r=0.6061, r^2 = 0.3674), meaning that almost 37% of the variation in performance was explained by predicted performance on other sizes.

This provides strong evidence, then, that good algorithms for authorship in one domain will also be good algorithms for authorship in other domains. In particular, we have hopes that a set of best practices established by looking at one particular data set will be a set of at least "good" practices in other domains, and can be a useful starting point in a search for domain-specific best practices in other, less studied or novel, domains. Unfortunately, the way that the AAAC data was structured prevents direct comparisons of accuracy (although it is hard to imagine ways to establish that two authorship attribution tasks are "comparably difficult" to enable such direct comparisons). Of course, at one level, one could make the argument that a bad algorithm for English should not be expected miraculously to perform better when transferred to a language the original designer cannot speak or read. But it is still encouraging to find that a good algorithm for English can be expected to perform well in that same unknown language.

References

Juola, Patrick. (2004). "Ad-Hoc Authorship Attribution Competition" ALLC/ACH 2004 Conference Abstracts. Gothenberg: University of Gothenberg.

Juola, Patrick. (2006/8). Authorship Attribution. Foundations and Trends in Information Retrieval 1:3. Delft:NOW Publishing.

Juola, Patrick, John Sofko, and Patrick Brennan. (2006). "A Prototype for Authorship Attribution Studies." *Literary and Linguistic Computing* 21:169-78

Appropriate Use Case Modeling for Humanities Documents

Aja Teehan
National University of Ireland
aja.teehan@nuim.ie

John G. Keating
National University of Ireland
john.keating@nuim.ie

1 Introduction

This paper argues that the most appropriate methodology to use when modeling a historical document for a software environment is one that focuses on modeling for functionality. This functionality is derived from Use Case modeling that should be undertaken in consultation with the User Group. The Use Cases are an expression of the ecological model as they embody the use of the document, by the User, in the software environment. The encoding mechanism largely practiced within the humanities computing community is represented by the TEI (Text Encoding Initiative)[1], which seeks to provide a set of guidelines for encoding humanities documents. However, TEI offers no guidance in relation to creating an encoding of a document that is supportive of the software environment that will host it, or the User. We argue that modeling with recourse to the Logical, the Physical and the Interaction classes enables not just the generation of an appropriate encoding scheme, but also the software to manipulate it. The argument is framed in relation to the creation of a digital edition of an 18th century Spanish Account Book manuscript.

2 Why Should Humanities Researchers Employ Use Case Modeling?

The humanities computing community currently lacks a formalised framework in relation to how to approach the task of modeling documents to reside in software environments. TEI (Text Encoding Initiative) focuses mainly on the skill of encoding using TEI, rather than the knowledge as to how best to encode, which is independent of any technical language. We would agree with the Kings College Humanities Computing course designers, who seeks to promote knowledge based, rather than skills based, learning[2]. It is much more diffcult, and more valuable, to create a good model than it is to encode with technical validity.

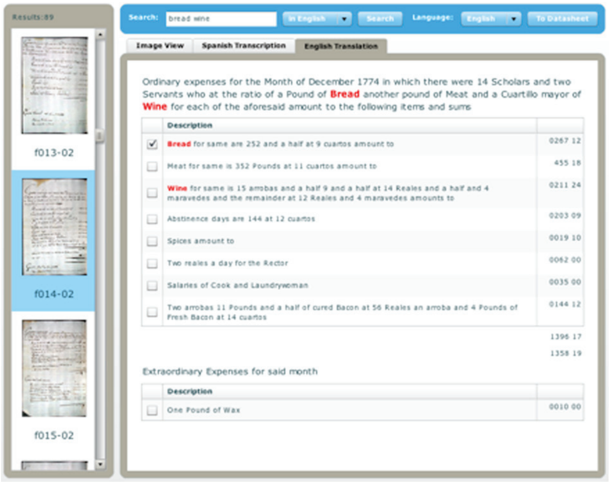
More and more the humanities researcher is involved in encoding their documents but remain uninvolved in software design. Unfortunately, this disjoin often results in clunky systems that are not used by the community. Bradley argues that ”HC might be more influential if it moved its operations closer to traditional scholarly methods”[3]. In order to ”be in a better position to develop a model of the role of computers that does more to support humanities research”[3] the encoder must consider the logical, physical and interaction classes in relation to their own domain when creating the Use Case and dependent encoding model. Modeling how the document is to be used, right down at the low level technical encoding, is of paramount importance.

3 Use Case Models

A Use Case acts as a blueprint for the system design and typically depicts the steps an actor takes while interacting with the software in order to achieve some meaningful goal or task, goal being higher level and task being lower level. These explicit steps are expressed in a formalised diagram using UML (Unified Modeling Language) and can be used by the software engineer to create a supportive software environment. This is a higher level of abstraction. The Use Cases model the ecological system, which is then used to build a software environment that encapsulate the functionality required by the researcher. At a much lower level of abstraction, the researcher who is involved in encoding their source must create a model of the source document that supports, and is part of, the ecological system model.

The Alcalá Project was originally proposed as a digital humanities project to mark a humanities collaboration between the University of Alcalá de Henares (UAH), Spain, and the National University of Ireland, Maynooth (NUIM), Ireland. The source was to be a Spanish eighteenth century account book recording the monthly expenses of the Royal Irish College of Saint George the Martyr. In eight weeks, the source manuscript was chosen, encoded and made available in a web based, dual language, searchable and interactive environment. More importantly, it was developed to aid the historian in answering historically pertinent research questions that are specifically prompted by the historical object, an account book. In the remainder of this paper we will refer to this digital artefact for examples. Please see figure 1.

In the digital edition of the Alcalá Account Book manuscript a transcription and translation are provided for the manuscript and are presented along with facsimile images of the original manuscript on a page-by-page basis. In addition, it is possible to transfer specific expenses to a datasheet for accounting operations to be performed upon them. This functionality was provided as a result of Use Case analysis. A typical example of the goal a user might wish to accomplish using the original manuscript might be, “calculate how much was spent on bread at the college in 1778”. At a lower level of abstraction, this Use Case requires six steps be performed by the User. If the User speaks only English then they must (1) select the translation as the version to search, (2) enter bread as the keyword for the search, (3) examine the returned facsimiles from 1778, (4) select the entries in the account book pertaining to bread using the checkbox, (5) transfer the pertinent entries to the datasheet for calculation, and (6) switch to the datasheet view to read the total.



Ordinary expenses for the Month of December 1774 in which there were 14 Scholars and two Servants who at the ratio of a Pound of **Bread** another pound of Meat and a Cuartillo mayor of **Wine** for each of the aforesaid amount to the following items and sums

	Description	
☑	**Bread** for same are 252 and a half at 9 cuartos amount to	0267 12
□	Meat for same is 352 Pounds at 11 cuartos amount to	455 18
□	**Wine** for same is 15 arrobas and a half 9 and a half at 14 Reales and a half and 4 maravedes and the remainder at 12 Reales and 4 maravedes amounts to	0211 24
□	Abstinence days are 144 at 12 cuartos	0203 09
□	Spices amount to	0019 10
□	Two reales a day for the Rector	0062 00
□	Salaries of Cook and Laundrywoman	0035 00
□	Two arrobas 11 Pounds and a half of cured Bacon at 56 Reales an arroba and 4 Pounds of Fresh Bacon at 14 cuartos	0144 12

1396 17

1358 19

Extraordinary Expenses for said month

	Description	
□	One Pound of Wax	0010 00

Figure 1: Bread and wine used as keyword filters, items of interest selected in English translation on one resultant page

4 A Model Framework: Logical, Physical and Interaction Classes

The above Use Case is a good example of how the software environment should support the User. However, there is no detail about how the steps should be achieved by the software, instead everything is from the User’s point of view. Additional steps must be performed by the software environment, for instance, the first step now becomes (1a) Interface presents translation, transcription and facsimile image of first page (1b) User selects translation as the version to search (1c) Interface presents translation on the screen. This poses a problem for the researcher charged with encoding the document.

The above Use Case requires information from three different classes: the logical, the physical and the interaction classes. The logical class is a model of the content of the document, e.g. monthly expenses; the physical class is a model of the document e.g. pages; the interaction class is a model of how the User interacts with the document and, by extension, how the User interacts with the software environment’s representation of that document.

The encoding, though only part of the ecological system model, must support the functionality in the Use Case and thus must support the three classes: logical, physical and interaction. Failure to support the interaction class will result in the software engineer being unable to fulfil the Use Cases. For instance, in the Alcalá Account Book encoding scheme each page was labelled with a unique page identifier, unlike the manuscript where only pages with text were marked by an archivist. This allowed the software engineer to return the exact page that a search requested. Without prior knowledge of the Use Case, "search for page number 10", and how the software engineer would implement that Use Case (the interaction class), and thus his technical requirements, this could not have been performed. Thus, the humanities researcher who encodes must also be aware of the interaction class and the requirements of the software engineer.

Furthermore, the encoding of the logical model must also take cognisance of the User's requirements. A single document can be researched in many different ways, for instance a historian may be interested in the social history captured in the Alcalá Account Book manuscript or they may be interested only in the prosopographical information that can be gleaned from it. Choices need to be made in representing the contents of the document. In relation to encoding specifically, tags must be created to give context to content, and segmentation of the document must be performed to decide what should be contextualised. These decisions should all be derived from the Use Case. While encoding the information the encoder must always be mindful that the Use Cases can be fulfilled. For instance, in the Alcalá Account Book encoding each of the expenses was labelled separately and broken down into its description and the sum spent. This allowed us to manipulate the figures separately on the datasheet so that mathematical operations could be performed. Without this separation of the sum spent we would have been unable to contextualise the figures as "money" and thus would have been unable to fulfill the Use Case, "how much was spent on bread in 1778?". Furthermore, without recourse to the interaction class component of this Use Case, the intuitive interaction offered by the clickable facsimile would have been foregone. In this alternative scenario, the figures required by the query are directly manipulated on the facsimile image. The User can click on those manuscript account book expenses that they wish to interact with. The expense items are simultaneously selected on the facsimile, in the translation text, and in the transcription text; they can then be sent to the datasheet for further manipulation. This simultaneous selection imparts to the User the sense that all the version are integrated, and are representative of the original encoding. The interaction becomes more intuitive, closer to the usability of the original document, but enhanced. Although it is both possible and necessary for a researcher skilled in humanities to be the primary articulator of Use Cases that encompass both the physical and logical classes, it is more diffcult to articulate those parts of the Use Case that derive from this interaction class and a dialogue should be opened here with a practitioner knowledgeable of Computer Science and Software Engineering.

5 Limitations of Use Cases

Use Cases have some well documented problems associated with them [4]. The most pertinent problem is that Use Cases can only be successfully used when the modeler has a full understanding of the problem domain, in this case, some humanities data or object. This limits their usefulness to cases when it is possible to fully understand the humanities data or object in question before the digitisation takes place. Data is always created in some context and can thus be understood. This is not necessarily true for humanities objects, such as novels. These are less definable and thus are sometimes digitised to aid in the investigation of their meaning. Use Cases are less useful when the aim of the digitisation is to promote prima facie discovery or investigation of the humanities object. To overcome problems associated with requirement drift it would also be important to combine this approach with an iterative design process, as opposed to a sequential. This would ensure that the Use Cases could also be updated to reflect the most current set of requirements for the project and help to avoid, "the biggest iteration of all, going back to the start"[5]. The Use Case is just one tool available to the humanities computing researcher from the arsenal of software engineering paradigms, for instance Rapid Application Development [6] and Participatory Design (where the end Users are actively involved as consultants in the design of the software ecological system) would be valuable [7]. Situating the design and use of Use Cases within these software engineering paradigms would be even more beneficial to the humanities computing community.

6 Are Use Cases Redundant?

Use Cases are sometimes considered to be the expression of the obvious through highly formalised means, the implication being that the administrative overhead incurred is not justified for the benefit that is produced. It may seem obvious to state that without identifying and then isolating the required pieces of information for a question, you cannot answer that question. However, this is a very basic, and very valuable, step that is missing from many digitisation projects. For instance, a digital repository of "The Chymistry of Isaac Newton" [8] offers diplomatic transcription, normalised version and correlated facsimile image for many documents, including Newton's most complete laboratory notebook. The documents are fully keyword searchable. The "ultimate goal is to provide complete annotations for each manuscript and comprehensive interactive tools for working with the texts"[9]. There is no doubt that this is a very valuable source. However, the encoding does not provide for the functionality that one would initially expect of such a collection, nor can this functionality be added later, without significant recoding. For instance, there is no support in the encoding for implementing a contextual search for logical model elements such as "experiment", "apparatus", "chemical", "method" or "conclusion". Instead the element tags are drawn from the prose, figures, linking, analysis, names/dates, and transcription tag sets of the TEI. Though already a very rich and rewarding source, it would benefit greatly from this type of functionality. Furthermore, once the Use Cases were elicited and described, the additional work involved in encoding this type of functionality would have been minimal. It may seem obvious to software engineering experts that to build a system one first has to define precisely what the requirements of that system are, but this is not the widespread practice within humanities computing.

7 Conclusion

Use Cases do not, of themselves, provide automatic quality and clarity of a digital artefact, or of the encoding built upon it. They function as a tool to aid the improvement of the ecological software environment so that the main requirements of the User can be satisfied. The creation and implementation of the Use Case still requires skill and knowledge, and still depends completely on the writer of the Use Case, the software engineer and the encoder. However, this level of knowledge would be demonstrably more valuable to humanities researchers than the text-encoding skills currently being promulgated. In relation to ascertainig the appropriate level of abstraction McCarty posed the question, "For us in the digital humanities, when and how does it matter that we know directly what's in the cellar?"[10] . We contend, "that from the outset (when it matters) researchers should know how, at least at a detailed pattern level, what they want to do, now and in the future (how it matters)." [11] This detailed pattern level is exemplified by the knowledge of how to create Use Cases. The researcher who wishes to operate at a lower level of abstraction and actually encode the humanities document must first have this high-level knowledge. In order to create an encoding scheme based around a document they should, in addition, have knowledge of the logical, physical and interaction classes. Only then can they appropriately apply that knowledge at the skill-level in an encoding. Both the problem domain (humanities research) and the software engineering pattern required to create appropriate Use Cases are very demanding of the researchers involved. Both areas demand high levels of expertise and understanding. Consequently, it is unusual to find this level of specialisation in one person. The solution is not to promote the assimilation of software engineering skills within humanities disciplines, but rather to promote the dialogue between the experts at a suitable design and abstraction level—that of the Use Case.

References

1. http://www..tei-c.org/ : Tei: P5 guidelines. Text Encoding Initiative. Available Online. Accessed 10 March, 2009.

2. Jessop, M.: Teaching, learning and research in final year humanities computing student projects. Literary and Linguistic Computing 20 (2005) 295–311

3. Bradley, J.: What you (fore)see is what you get: Thinking about usage paradigms for computer assisted text analysis. Text Technology 14 (2005) 1–18

4. Firesmith, D.G.: Use case modeling guidelines. Technology of Object-Oriented Languages, International Conference on (1999) 184

5. Dominick, P.G.: Tools and Tactics of Design. Wiley & Sons (2000)

6. Martin, J.: Rapid Application Development. New York: Macmillan (1991)

7. http://www.cpsr.org/issues/pd/ : Participatory design. Computer Professionals for Social Responsibility. Available Online. (2008) Accessed 10 March, 2009.

8. http://www.dlib.indiana.edu/collections/newton: Newton, Issac. "Newton's Most Complete Laboratory Notebook". The Chymistry of Isaac Newton. Ed. Newman, W.R. 11 February 2006. Available online. Accessed 10 March, 2009.

9. http://www.dlib.indiana.edu/collections/newton : The Chymistry of Isaac Newton. Ed. Newman, W.R. 11 February 2006. Available online. Accessed 10 March, 2009.

10. McCarty, W.: Signs of times present and future. Available Online. Humanist Discussion Group Vol. 22, No. 218 (2008) Accessed 10 March, 2009.

11. Keating, J.: Signs of times present and future. Available Online. Humanist Discussion Group Vol. 22, No. 219 (2008) Accessed 10 March, 2009.

Delivering a Humanities Computing Module at Undergraduate Level: A Case Study

John G. Keating
National University of Ireland
john.keating@nuim.ie

Aja Teehan
National University of Ireland
aja.teehan@nuim.ie

Thomas Byrne
National University of Ireland
thomas.l.byrne@nuim.ie

1 Introduction

In September 2008 we commenced delivery of a Humanities Computing module at undergraduate level to twenty seven students of the National University of Ireland, Maynooth (NUIM). The module was offered by the Computer Science Department and operated and delivered by An Foras Feasa and was designed and delivered, pedagogically, as an *authentic learning experience* [1]. This paper gives an overview of the aims of the module, the curriculum, the module assignment and student support, assessment, and the process for evaluating the module. We also provide details on the continuous assessment project based on the creation of a repository holding a collection of digitised 19th and 20th century religious pamphlets.

2 Humanities Computing Module

A key aim of the module is to foster in the students some understanding of humanities problems, to recognise how and when a solution could be found using Computer Science Software Engineering (CSSE) methods, and finally, how to apply CSSE methodologies and technologies to create solutions. Having researched the King's College London (KCL) Humanities Computing final year project [2] offered by the Centre for Computing in the Humanities (CCH), we decided that an emphasis should be placed on the general application of CSSE principles rather than specific tools, which can date quickly. Figure 1 provides the for the module description as it appeared in the NUIM's Book of Modules.

The module was delivered over 12 two-hour lectures, supplemented by 12 two-hour labs. The student was expected to perform and additional 32 hours of independent study. The lectures provided the theoretical framework for, and overview of the issues associated with, humanities computing. The labs were considered to be of paramount importance and a strong emphasis was placed on the delivery of a digital artefact by the end of the semester.

The module has two primary learning objectives: firstly, the students would know how to design and create a digital artefact to solve a humanities problem, and, secondly, they would hold informed opinions as to why specific technologies and methodologies were chosen. In this way one can see that the methodologies of the two main disciplines (Humanities and CSSE) were used as the framework for the students' evaluation and critique.

3 Module Assignment and Student Support

We applied a pedagogical approach known as authentic learning in the design of the module; this approach "allows students to explore, discuss, and meaningfully construct concepts and relationships in contexts that involve real-world problems and projects that are relevant to the learner" [1]. The continuous assessment assignment was chosen because it represented a typical real-world, humanities researcher, problem. The form of the continuous assessment assignment would mean that the students would have to engage, in a real-world sense, with the assigned problem.

The continuous assessment assignment was based upon creating a digital artefact derived from 19th and 20th century pamphlets produced by the Catholic Church in Ireland. The assignment was to be completed within the laboratory and independent study times. Each two-hour lab session was presided over by a demonstrator who was available to answer queries in relation to the work. While the lecturers would provide guidance, instruction and information in relation to the practical project, it was also expected that the students would use outside resources to research and implement their system. The assignment was formalised as in Figure 2.

It was also decided that the course assignment would be conducted by teams of students. Team work is often required in industry thus this learning experience, beneficial in itself, would aid the team-members in their career. Team work is also a necessity in most software engineering projects as they generally require diverse and wide-ranging expertise unlikely to be found in one individual. Each team had at least one student whose home discipline was Computer Science and Software Engineering (CSSE).

It was considered important to provide the students with an understanding of the technology most used by the Humanities Computing community, therefore, TEI (Text Encoding Initiative) was selected as the encoding language rather than XML (Extensible Mark-up Language). However, of the time spent in lectures on encoding, roughly half of that was devoted to XML in order to provide both context for TEI and a basic working knowledge of encoding. The students were expected to evaluate the use of TEI as the encoding language in their project presentation and in their written assessment.

The humanities contextualisation derived from the need to understand the problem domain of the User. Early lectures provided an overview of the course content, expectations of the students, examination system, etc. The remainder of the lectures covered the module content as described above. The authentic learning experience was supported in the lectures by inviting the post-doctoral fellow as a guest speaker. This allowed for the framing of the humanities problem (mining 19th and 20th century pamphlets for icons and related text) for the students. The course team believed that the students would be motivated to create the digital artefact knowing that its value would be evidenced in the real-world; the students would "be able to realize that their achievements stretched beyond the walls of the classroom" [3].

4 Assessment

The balance between *doing* and *evaluating* was reflected in the marking scheme of the module; 50% of the marks were available for continuous assessment deriving from the lab-supervised production of the digital artefact. The remaining 50% came from a written examination at the end of the semester.

The written examination was open book and consisted of three questions. The students chose two to complete. Each question had three sections: the first section required basic knowledge of some topic in humanities computing, the second section required the implementation of a solution to a given problem within that topic, and the third section required evaluation and critique of the chosen technology and methodology. In order to pass the assessment a student had to display at least an understanding of the topic and some form of critical thinking or solution generation.

The authentic learning approach taken in the design of the module equates to the KCL final year project model in many respects. However, the evaluation of the finished digital product only formed part of the team presentation, which accounted for 5% of the overall marks, or 10% of the continuous assessment marks. It was during this presentation that students also had the opportunity to display the internalisation of the issues surrounding humanities computing and their acquisition of the associated language register. This assessment was particularly suited to evaluating how successful the authentic learning approach had been.

However, the authentic learning approach also dictated that the assessment should be based on real-world concerns. In computer science, the delivery of a functional digital product, before the deadline and to an acceptable standard is the fundamental requirement. While KCL stresses that the project can still be a *success* even if the technological aspect was a failure, we stressed that the technological fulfillment of criteria was of critical importance. Failure to deliver a digital artefact, for any reason and no matter how well evaluated, would have resulted in failure in the continuous assessment, thereby making it diffcult for the students to pass the module.

The evaluation of the humanities issues and related technologies was to be mainly expressed in the written examination, which equated to 50% of the marks. In this way we sought to balance the module, just as KCL balanced their final project.

5 Evaluation

Evaluation of the module, which will be available in advance of the DH2009 conference, will be undertaken in two ways. Firstly, a post-doctoral research fellow will actually use the repository. Their opinion on how successful the repository is in fulfilling their needs will be highly valued. The students' understanding of the humanities and computer science issues discussed during the lectures will also be displayed in their written examinations. We are optimistic about their success in relation to the creation of the repository as already we have seen evidence of external research and the use of previously unfamiliar software tools to create considered Project Plans and Use Cases. Secondly, a detailed survey of the students will be undertaken where the students will be asked to critique the module.

6 Conclusion

Designing the module using an authentic learning approach has provided the students with the opportunity to learn very important life skills in relation to team work, deadlines, scarcity of resources and independent research. This has been balanced by providing the students with a theoretical and methodological framework so that an understanding of the problems in the humanities domain that can benefit from computing has been fostered. It is hoped that this module will help, within the university, to foster a greater understanding of, and appreciation for, humanities computing.

References

1. Suzanne M. Donovan, John D. Bransford, J.W. Pellegrino, ed.: How People Learn: Bridging Research and Practice. National Academy Press, Washington DC (1999)

2. Jessop, M.: Teaching, learning and research in final year humanities computing student projects. *Literary and Linguistic Computing* 20 (2005) 295–311

3. Mims, C.: Authentic learning: A practical introduction & guide for implementation. Meridian 6 (2003) 1–3

C2260 HUMANITIES COMPUTING 1

24 Lecture hours, 24 Laboratory hours, 32 Tutorial and Independent Study hours. Lectures, Tutorials and Laboratories will be delivered by An Foras Feasa staff.

ECTS Credits: 5.0
Learning Outcomes: To understand and become familiar with technology for the humanities. To examine how computing tools and techniques may be integrated into humanities and what the effects of this integration might be. Students will learn how to use technology to inform new humanities research as well as how it may be used to support existing research patterns;
possibilities created by the application of technology to humanities research can be explored and exploited. Students will learn how to apply a wide variety of computing and software engineering techniques in four distinct areas: text processing, image processing, software engineering and digital
humanities. How to assess whether computing can be usefully applied in particular circumstances, and what results may be expected.

Content: Fundamentals of the Digital Humanities: humanities research, electronic communications and publishing, text analysis, numerical and graphical analysis and presentation. Introduction to web-based databases. Analysing text: electronic tools to analyse written and transcribed text. What can be learned about a text from using simple analytical techniques? Using graphical analysis to find and understand patterns in data. Creating and using graphs and charts to summarise, visualise and analyse data. The nature of images and how to convert them to digital form: from manuscript, through image capture and XML, to online database. Digital images; commonly used
operations and techniques for image enhancement. Digitisation and the creation of online digital resources. Introduction to TEI (Text Encoding Initiative), XML (Extensible Markup Language) and associated tools. Assessment: Total Marks: 100%. Two-hour written examination at the end of Semester 1, not less than 50%. Continuous Assessment up to 50%

Figure 1. CS260 Module Description

CS260: HUMANITIES COMPUTING 1
Laboratory Project
John Keating, Aja Teehan, An Foras Feasa

You are required, as part of a four or five person team, to build an on-line repository of 18th and 19th century pamphlets. The repository will include both image and fulltext encoding of the pamphlet. The text encoding standard used will be TEI (Text Encoding Initiative), a comprehensive text encoding methodology.

You will be provided with a collection of preservation-quality images for several related pamphlets. These provided images must be encoded using TEI. Furthermore, each team is expected to use the imaging facilities in An Foras Feasa's laboratory to capture at least two additional pamphlets (approximately 2 hrs. of imaging per team).

The team is expected to produce an on-line user interface to access the digital images and associated text. The team should implement interaction facilities for at least two Use Cases (for example, "search the repository for occurrences of keywords", or "implement linkages between pamphlets").
Full-text and image TEI encoding (35%)

Project will be graded on 5 deliverables:

1. Production of the on-line repository, including User Interface (35%)
2. The team's preservation-quality image collection (5%)
3. Project presentation by the team (15 minutes) (10%)
4. Full project documentation (15%)

Key dates for deliverables:

- Project plan submitted by 10th October (formative assessment)
- Use Cases diagrams/documentation submitted by the 24th October (formative assessment)
- Encodings submitted by 14th of November (formative assessment)
- Project presentation on the 15th and 16th of December (summative assessment)
- Project documentation submitted by 19th of December (summative assessment)

Figure 2. Continuous Assessment Assignment

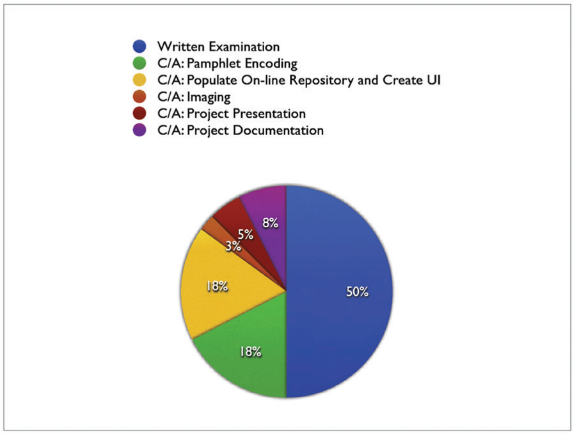


Figure 3. Marking Scheme for CS260 Module

MIHS Text Mining Historical Sources using Factoid

Sharon Webb
National University of Ireland
sharon.webb@nuim.ie

John G. Keating
National University of Ireland
john.keating@nuim.ie

1 Introduction

This paper provides an overview of methods used in bespoke software, MIHS (Mining Interactive Historical Sources), currently under development. This paper considers associational culture and the development of Irish nationalism through the anthropological idea of *Othering* [1] and how customised software aids knowledge construction and historical research.

A brief overview of the historical question reveals that the concept of the Other is a changing manifestation of socio-economic and political conditions. It constructs group identity by labeling and categorising the core group but more importantly those outside it. Irish identity and nationalism, for example, during the eighteenth century is immersed in ideas of Othering, as political and social expressions, such as the Penal Laws[1], demonstrate how one group, Protestants, ascribe negative attributes to label and define the Other group, Catholics, as second class citizens. This historical research uses customised software that aids knowledge construction by providing an interface and central environment to digitise, store and share sources. The software provides an application to create and extract information by developing *factoids*, defined as the connection of "different kinds of structured information" [2], from primary and secondary sources. These factoids are dynamically generated by the system in response to researcher-defined queries. They are composed of *factlets* and additional XML encoded source information; factlets themselves are observations on sources, chosen and manually encoded by the researcher. The environment also uses data mining, or knowledge mining [3], techniques to generate organised clusters of factoids, called *data clusters*.

MIHS provides a user friendly interface for historians to create and generate an information environment driven by their research question. It creates a database containing relevant information about sources so, for example, a bibliography and footnotes can be produced. However, the database is "an intermediate act, not a final one" [2]. The software supports the user in the generation of factoids, the research-driven encodings of researcher-relevant information, derived from uploaded images of primary and secondary sources. It therefore allows the user to refine and define information of interest and value in order to solve, or inform, the research question. The generation of factoids encapsulates the researchers' thinking and demonstrates links and relationships between sources and theory based on the research question.

2 Historical Research using Factoids

Bradley and Short discuss the use of factoids in relation to prosopography as a means to provide material for future research and assert that "a collection of factoids does not record a 'scholarly overview' of a person [event] that a scholar has derived from the sources s/he has read" [2]. In this context the factoids represent pure information and are not driven or shaped by any research question or agenda. Projects such as *The Prospography of the Byzantine Empire*, described by Bradley and Short, provide a vast array of information searchable by, amongst others, factoid type and source type, and are invaluable to anyone interested in this field of study. The use of factoids ensures information is presented in a structured, relation-based fashion. It provides interaction between different sources and enhances scholarly research. However, Bradley and Short argue that "one of the diffculties with a factoid approach is to establish what kind of 'factoids' should be collected from the source" [2]. With this approach, those creating the prosopography database must predict what future users may or may not require.

In contrast, *MIHS* is research driven – factoids are produced in the context of the research question. The software is not only a digital research tool but forms an integral part of the methodology as the software creates factoids based on a "scholarly overview" [2] and scholarly interpretation. The source is encoded using XML (see Figure 1,2,3), though an abstraction of XML is presented to the user where element names and attributes are defined. This ensures the user controls the schema and subsequent representations of the source within the software (tags and categories previously used will be highlighted to ensure continuity) and the actual "task of categorising, grouping and ordering" [2] sources opens a conversation between the user, the sources and the research question.

```
<factlet id = ”SS0012-F0001” title = ”Con-
testing Ireland”>
     <category> Penal Laws </category>
    <category>Catholic identity</category>
    <text>
    The laws helped fashion an
    <key> identity for Catholics</key>
```

```
        who in fact had developed differences
        among themselves over the centuries of
        <key>colonialism</key> ...
        </text>
        <source id = ”SS0012” imageid=”SS00012.
        jpg” x=”200” y=”197”>
    </factlet>
```

Figure 1: Sample XML of a factlet derived from a source

```
    <source id = “SS0012” type = “secondary”
    imageid = “SS00012.jpg”
        <title>
           Contesting Ireland, Irish voices
           against English in the 18th century
        </title>
        <author> Thomas McLoughlin </author>
        <date> 1999 </date>
        <publisher> Four Courts Press Ltd. </
        publisher>
        <location> Dublin </location>
        <isbn> 1851824480 </isbn>
    </source>
```

Figure 2: Sample XML detailing source information linked to Figure 1

```
    <factoid id = Penal Laws>
        <factlet id = SS0012-F0001/>
        <factlet id = PS0001-F0002/>
        <text date = 09.09.08>
    The <link> Penal Laws </link> were used to
    alienate a whole people; yet by doing this
    the <link> Protestant elite</link> ensured
    that <link>Irish Catholics</link> formed
    and solidify an identity rooted in the ex-
    clusive nature of the laws.
        </text>
        <text date = 15.09.08>
        Associative groups such as the
    <link>Catholic Committee</link> develop
    and articulate the Catholic voice through
    protest against these laws. Instead of
    disabling the Catholic majority the laws
    provide an important target of organised
    protest.
        </text>
        <text date = 25.09.08>
                (Shows the contradictions of
                the <link>Age of Enlighten
                ment</link>
        </text>
    </factoid>
```

Figure 3: Sample XML detailing the construction of a factoid

Factoids are produced through user and source interaction and the utilisation of the software. To create factoids the user must first define numerous factlets from a source. Each factlet inherits core information from the original source such as location, date, author, etc. Factlets from different sources that are related by subject or category, for instance, then merge to create a factoid. The user will extract information from encoded sources whilst preserving the integrity of the information through easy access to the original source image, thus maintaining source context. Metadata, providing original source information, is available for each factlet contained in a factoid (and indeed factoids presented in data clusters) and is expressed using the Dublin Core Metadata standard.

A typical use, for example is as follows “How do the Penal Laws categorise Catholics as the Other in Irish society?” To answer this question, pamphlets, periodicals, newspapers and other text based sources are used. Figure 4 demonstrates the interaction between source material and the creation of factoids. This example is constructed from a small dataset. As the number of sources increase the benefit of this type of data organisation becomes more obvious. The factoid in Figure 4 represents sources which are concerned with the Penal Laws. In a larger dataset, this factoid can be refined or defined by using relational operators and set relations, creating factoids specific to the user.

Another project that has yielded a database system for prosopography, the COEL database, states a fundamental approach requires that “the data must always be protected against contamination by the interpretation” [4]. This type of approach may be required when presenting historical sources and information for general consumption but in *MIHS*, user interpretations are paramount. *MIHS* presents users arguments and reflects his/her line of thought through the construction of factoids. The ability to return to original sources ensures context remains and leaves a forum open for historical debate.

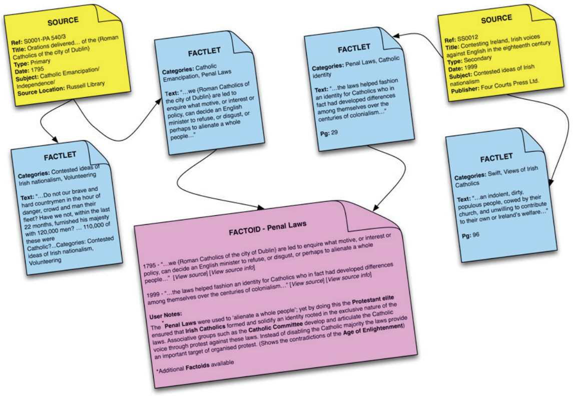

SOURCE
Ref: S0001-PA 540/3
Title: Orations delivered… of the {Roman Catholics of the city of Dublin}
Type: Primary
Date: 1795
Subject: Catholic Emancipation/ Independence/
Source Location: Russell Library
FACTLET
Categories: Catholic Emancipation, Penal Laws
Text: "…we {Roman Catholics of the city of Dublin} are led to enquire what motive, or interest or policy, can decide an English minister to refuse, or disgust, or perhaps to alienate a whole people…"
FACTLET
Categories: Penal Laws, Catholic identity
Text: "…the laws helped fashion an identity for Catholics who in fact had developed differences among themselves over the centuries of colonialism…"
Pg: 29
SOURCE
Ref: SS0012
Title: Contesting Ireland, Irish voices against English in the eighteenth century
Type: Secondary
Date: 1999
Subject: Contested ideas of Irish nationalism
Publisher: Four Courts Press Ltd.
FACTLET
Categories: Contested ideas of Irish nationalism, Volunteering
Text: "…Do not our brave and hard countrymen in the hour of danger, crowd and man their fleet? Have we not, within the last 22 months, furnished his majesty with 120,000 men? … 110,000 of these were Catholic?...Categories: Contested ideas of Irish nationalism, Volunteering
FACTLET
Categories: Swift, Views of Irish Catholics
Text: "…an indolent, dirty, populous people, cowed by their church, and unwilling to contribute to their own or Ireland's welfare…"
Pg: 96
FACTOID - Penal Laws
1795 - "…we {Roman Catholics of the city of Dublin} are led to enquire what motive, or interest or policy, can decide an English minister to refuse, or disgust, or perhaps to alienate a whole people…" [View source] [View source info]
1999 - "…the laws helped fashion an identity for Catholics who in fact had developed differences among themselves over the centuries of colonialism…" [View source] [View source info]
User Notes:
The *Penal Laws were used to 'alienate a whole people'; yet by doing this the Protestant elite ensured that Irish Catholics formed and solidify an identity rooted in the exclusive nature of the laws. Associative groups such as the Catholic Committee develop and articulate the Catholic voice through protest against these laws. Instead of disabling the Catholic majority the laws provide an important target of organised protest. (Shows the contradictions of the Age of Enlightenment)
*Additional Factoids available

Figure 4: Factlets from sources, merged to form a Penal Law factoid

The production of factoids is followed by the use of data mining techniques to generate data clusters. This approach will help with the complexities and magnitude of large data sets and the presentation of large quantities of factoids.

3 Text mining using factoids in MIHS

Data mining methodologies are commonly used within retail, marketing and financial industries. They are used to derive patterns, associations and correlations from small to large datasets, utilising database architecture such as data warehousing, which result in the extraction of data knowledge [3]. Data mining adds value to raw data. Yet apart from its advantages in terms of statistics, data mining techniques can be used for text documents – text mining. Like data mining, it derives patterns and relations from texts and can be used "as partof-speech tagging, word sense disambiguation and bilingual dictionary creation" [5]. It can also be used to create synopsis of text by (excluding -the, of, is etc.) calculating frequently used words.

MIHS will use text mining to provide graphical models of data where factoids are represented as nodes on a tree. By using text mining we can extrapolate extra information either from the original text and/or related factoids. Text mining techniques enable the software to carry out text analysis on source material. For instance, the researcher can input milestones like the 1798 Rebellion, and, through language comparison and analysis, interpret changes in society. This may allow the researcher to move towards answering questions such as "After the failure of the 1798 Rebellion is there evidence of a change in attitude towards Irish Catholics?" This technique often yields interesting results because "even if a word only appears once or twice it can be significant if it does not appear at all in the [text] used for comparison" [6]. By mining factoids related to the Penal Laws and specifying dates such as 1798 – which has important consequences for both Irish Protestants and Catholics – the changing socio dynamics emerge through text comparisons, highlighting changes in descriptions and attitudes towards and within the different religious groups.

4 Conclusion

The ultimate aim of the software is to create an on-line community of historians, research and sources, supporting individual and collaborative projects. It will facilitate access to sources, facts and factoids related to research projects, which can be viewed as separate entities or part of scholarly interpretation. *MIHS* will provide a generic information platform for historians, moving away from software designed for specific research projects, to create a platform for historical debate where users can share sources, factoids and, of course, ideas. *MIHS* will allow for the construction of database architectures without the complexities often inherent in the creation of historical databases.

The process of managing the vast array of sources required for historical research can be a mammoth task, both in terms of handling the large volumes of data and the subsequent interpretation of sources. *MIHS*, through the creation of factoids and use of text mining, provides the means to store and organise sources collected. By allowing for self organisation of data, recommendations for data mining, references to related data and manuscripts, the creation of factlets and factoids, among others, *MIHS* will serve as an important tool in creating well-read and well-informed historical projects.

Notes

[1] The Penal Laws were a series of anti-Catholic laws passed in Ireland during the seventeenth and eighteenth century.

References

1. McGrane, B.: Beyond Antropology, Society and the Other. Columbia University Press (1989)

2. Bradley, J., Short, H.: Texts into databases: The evolving field of new-style prosopography. Literary and linguistic computing **20** (2005) 3–24

3. C.R. Rao, E.W., Solka, J., eds.: Handbook of statistics 24, Data mining and data visualiztion. Elservier (2005)

4. Keats-Rohan, K.: Historical text archives and prosopography: the coel database system. History and Computing **10** (1998) 57–72

5. Hearst, M.A.: Untangling text data mining. http://people.ischool.berkeley.edu/ hearst/papers/ac199/ac199-tdm.html

6. Welling, G.: Can computers help us read history better? computerised text-analysis of four editions of the outline of american history. History and Computing **13 (2)** (2004) 151–160

Paraphrase Learning in Two Phases For Steganographic Communication

Katia Lida Kermanidis
Ionian University
kerman@ionio.gr

Introduction

Given an original sentence, that conveys a specific meaning, paraphrasing means expressing the same meaning using a different set of words or a different syntactic structure. Paraphrasing has been used extensively for educational purposes in language learning, as well as in several NLP tasks like text summarization (Brockett and Dolan, 2005), question answering (Duclaye et al., 2003) and natural language generation. Recently it has found yet another use in steganography.

Regarding paraphrase identification and generation, previous approaches have utilized supervised (Kozareva and Montoyo, 2006) or unsupervised (Barzilay and Lee, 2003) machine learning tech-niques, finite state automata (Pang et al., 2003), syntactic dependency rules (Meral et al., 2007), statistical machine translation techniques (Quirk et al., 2004).

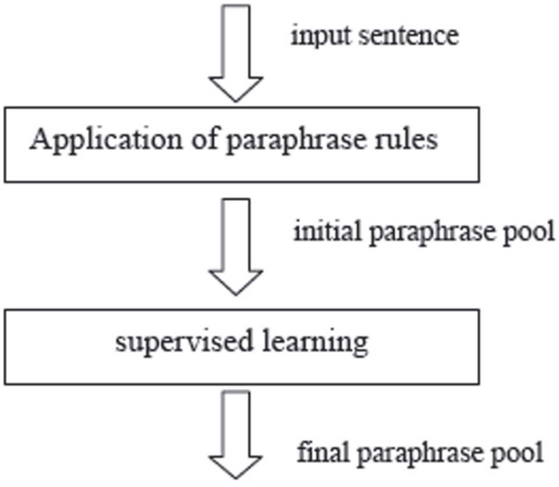


Figure 1. Paraphrase learning in two phases.

In the present proposal, paraphrases of Modern Greek free text are learned in two phases. Henceforth, the term "paraphrasing" will stand for shallow syntactic transformations, i.e. swaps of consecutive phrasal chunks. Modern Greek is quite suitable for shallow paraphrasing, due to the permissible freedom in the ordering of the phrases in a sentence.

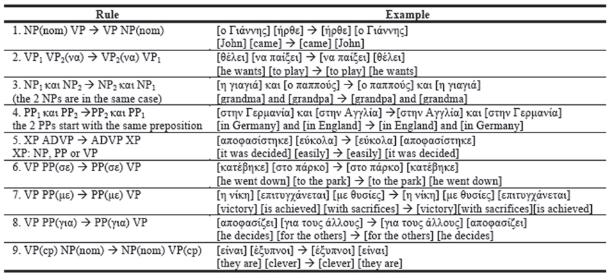
The paraphrase learning process is based on resource economy: the desire to utilize as minimal linguistic resources as possible, enabling thereby the methodology to be easily applicable to other morphologically rich languages like Modern Greek.

The paraphrased text may then be used for hiding secret information. Steganographic security will depend on the correctness and the naturalness of the paraphrases. Figure 1 shows the architecture of the paraphrase learning process.

Phase 1: The paraphrasing rules

The text corpus used in the experiments is the ILSP/ELEFTHEROTYPIA corpus (Hatzigeorgiu et al., 2000). It consists of 5244 sentences; it is balanced and manually annotated with morphological information. Further (phrase structure) information is obtained automatically by the chunker described in detail in (Stamatatos et al., 2000). During chunking, noun (NP), verb (VP), prepositional (PP), adverbial phrases (ADP) and conjunctions (CON) are detected via multi-pass parsing. The chunker exploits minimal linguistic resources. Phrases are non-overlapping.

A set of nine empirical bidirectional rules is first applied to the input sentences in order to change their phrase ordering. The complete set of rules is described in detail in table 1. Unlike the syntactic tools presented in (Meral et al., 2007), that may be applied only once to a given sentence, each of the rules described here may be applied multiple times (i.e. in multiple positions) to a sentence. Furthermore, more than one rules may be applied to a sentence simultaneously.



Rule	Example
1. NP(nom) VP → VP NP(nom)	[ο Γιάννης] [ήρθε] → [ήρθε] [ο Γιάννης] [John] [came] → [came] [John]
2. VP$_1$ VP$_2$(να) → VP$_2$(να) VP$_1$	[θέλει] [να παίξει] → [να παίξει] [θέλει] [he wants] [to play] → [to play] [he wants]
3. NP$_1$ και NP$_2$ → NP$_2$ και NP$_1$ (the 2 NPs are in the same case)	[η γιαγιά] και [ο παππούς] → [ο παππούς] και [η γιαγιά] [grandma] and [grandpa] → [grandpa] and [grandma]
4. PP$_1$ και PP$_2$ →PP$_2$ και PP$_1$ the 2 PPs start with the same preposition	[στην Γερμανία] και [στην Αγγλία] →[στην Αγγλία] και [στην Γερμανία] [in Germany] and [in England] → [in England] and [in Germany]
5. XP ADVP → ADVP XP XP: NP, PP or VP	[αποφασίστηκε] [εύκολα] → [εύκολα] [αποφασίστηκε] [it was decided] [easily] → [easily] [it was decided]
6. VP PP(σε) → PP(σε) VP	[κατέβηκε] [στο πάρκο] → [στο πάρκο] [κατέβηκε] [he went down] [to the park] → [to the park] [he went down]
7. VP PP(με) → PP(με) VP	[η νίκη] [επιτυγχάνεται] [με θυσίες] → [η νίκη] [με θυσίες] [επιτυγχάνεται] [victory] [is achieved] [with sacrifices] → [victory][with sacrifices][is achieved]
8. VP PP(για) → PP(για) VP	[αποφασίζει] [για τους άλλους] → [για τους άλλους] [αποφασίζει] [he decides] [for the others] → [for the others] [he decides]
9. VP(cp) NP(nom) → NP(nom) VP(cp)	[είναι] [έξυπνοι] → [έξυπνοι] [είναι] [they are] [clever] → [clever] [they are]

Table 1. The set of paraphrasing rules.

For every sentence all the possible combinations of rule applications are formed. This is the initial pool of paraphrases and in the given corpus its size may vary from zero (the sentence does not allow for any paraphrasing) to 80 paraphrases.

Figure 2 shows the distribution of the number of sentences depending on the sentence length (i.e. the number of chunks forming the sentence). Figure 3 shows the distribution of the number of sentences depending on the initial paraphrase pool size. Almost 80% of the sentences have at least one paraphrase, an impressive number, given that more than 24% of the input sentences consist of five or less chunks.

Phase 2: Paraphrase learning

Due to the use of the paraphrased sentences in steganography, correctness in syntax as well as naturalness is of great significance. Steganographic security depends largely on paraphrasing accuracy. Therefore the produced paraphrases are further filtered using supervised learning.

The positions of possible phrase swaps in the input sentences are identified, according to the nine paraphrasing rules. A learning vector is created for each original input sentence and each swap position. The features forming the vector encode morphological and syntactic information for the phrase right before the swap position, as well as two phrases to the left and two phrases to the right. Thereby, context information is taken into account. Each phrase is represented through a set of six features, shown in table 2. The morph feature denotes whether a noun phrase contains a definite or indefinite article and its grammatical case. The num feature is the number of tokens that constitute the phrase.

Table 2. The features of the learning vector.

	1	2	3	4	5	6
NP	NP	case	morph	presence and type of a pronoun	presence of a genitive element	-
VP	VP	-	-	word introducing the phrase	copularity	-
PP	PP	-	-	preposition introducing the phrase	-	-
CON	CON	-	-	lemma of the conjunction	-	num
ADP	ADP	-	-	lemma of the adverb	-	num

A total of 5 x 6 features constitute the feature vector, plus the binary target class: valid (yes) / not valid (no) paraphrase. Native speakers have manually annotated 519 instances with the correct class label. 26.4% of them are classified as incorrect paraphrases.

The following table shows the prediction results for various stand-alone classification algorithms: decision trees (unpruned C4.5 tree), k-NN instance-based learning (k=5), support vector machines (first degree polynomial kernel function, sequential minimal optimization algorithm for training the classifier). Accuracy is the number of correctly classified instances divided by the total number of instances. Experiments were performed using 10-fold cross validation.

	C4.5	k-NN	SVM
Accuracy	75.9%	77.9%	79.2%

Table 3. Results for stand-alone classifiers.

	Bagging	Boosting
Accuracy	80.3%	80.1%

Table 4. Results for ensemble learning schemata.

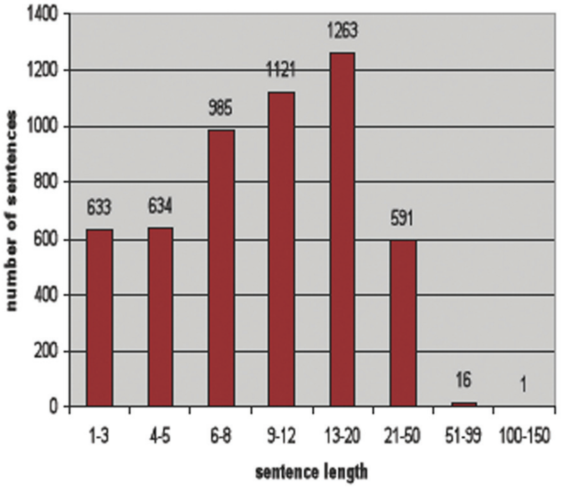


Figure 2. Distribution of sentence length.

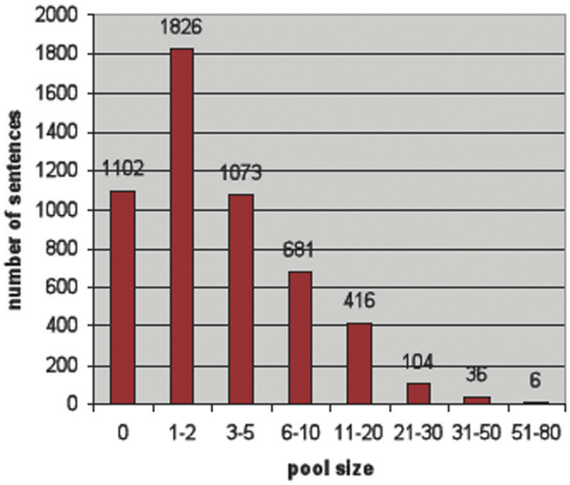


Figure 3. Distribution of paraphrase pool size.

The majority of the incorrectly classified instances are negative (not valid) instances, probably due to their rare occurrence in the data, compared to the positive instances. Support vector machines deal better with predicting negative labels, and reach an f-score of 64.1% for the rare class.

To improve classification accuracy even further, bagging and boosting have also been experimented with. The C4.5 unpruned classifier was used as a base learner for bagging (the optimal bag size was 50% of the training set and 10 iterations were performed) and boosting (Ada-

Boost, again 10 iterations were performed). Bagging leads to the best f-score for the negative class: 65.3%.

The positively labeled paraphrases from the previous phase constitute one part of the final pool of paraphrases. This part consists of paraphrases that have been produced by single phrase swaps, and not by combinations of swaps. The other part (due to the fact that the learning process does not allow for a paraphrase to be formed by combinations of phrase swaps) is formed by those paraphrases derived from phase 1 that are combinations of two or more correct phrase swaps (the positively labeled individual phrase swaps defined by phase 2, i.e. the first part of the final pool)

Application to steganography

Steganography is the art of embedding hidden information in unremarkable cover media in a way that does not arouse an eavesdropper's suspicion to the existence of hidden content underneath the surface message (Provos and Honeyman, 2003; Atallah et al., 2000; Topkara et al., 2005).

Once the final pool of paraphrases is formed for every sentence in the input (cover) text, the steganographic process starts. A secret message, i.e. a sequence of bits, is to be hidden underneath the cover text.

First, each rule is marked with one bit value, depending on its *condition*. By *condition* we mean the right or the left-hand side of the rule (right or left-hand side of the arrow in Table 1). For example, for Rule 1 a bit value "0" could mean the left hand side of the rule, and then a bit value "1' would indicate its right-hand side. In the case of symmetrical rules (Rules 3 and 4), the condition may be determined by considering as NP1 the noun phrase which starts with a letter closer to the beginning of the alphabet that NP2. This rule marking results from a prior understanding between the communicating parties.

The embedding process is then completed in two stages. In a first stage, for every sentence, a paraphrase is selected from its pool. The selection may be performed in a round-robin fashion (i.e. to choose the paraphrase of each rule one at a time), or based on a secret (e.g. a symmetric cryptographic key) shared between the two communicating parties. In case the size of the pool is zero, the sentence remains unchanged, and it is not used for information embedding. If, however, the pool size is greater than zero, a selection is possible and the sentence is useful for information embedding. In the second stage, depending on the condition of the selected rule, a secret bit is embedded as follows: if the bit to be hidden is the same as the condition of the rule, the rule is not applied and the sentence remains unchanged, otherwise it is applied and the sentence is paraphrased. For example, a subject-verb sequence in the input sentence would mean a condition "0" for Rule 1. If the bit to be hidden is also "0", Rule 1 is not applied and the sentence is transmitted as it is. If the hidden bit were "1", the rule is applied and the sequence in the transmitted sentence now reads verb-subject, instead of subject-verb.

On the other end, the extractor receives the final text. Having at his disposal the same rule set, (s)he is able to identify the rules that may be applied to each sentence. Sharing the same secret key used in the embedding process, (s)he is able to select the same rule used in the insertion process. For example, reading a subject-verb sequence, and knowing that this sequence indicates a bit value "0" for the condition of Rule 1, (s)he decides on "0" to be the first secret bit. Reading a verb-subject sequence would have meant a condition value "1" and (s) he would have decided on "1" to be the first secret bit.

Steganographic Capacity

To obtain a feeling of steganographic capacity, assuming an average word size of 6 bytes/word, and given that our corpus consists of 166.000 words, the corpus size equals roughly 1 Million bytes. Steganographic capacity (the available bandwidth) may be evaluated as follows: Using the current implementation which allows for the embedding of one bit per paraphraseable sentence, 4142 secret bits may be embedded in the corpus. In other words, 1 bit may be embedded every 2000 bits of cover medium size. This bandwidth may increase by exploiting the possibility of embedding more than one bits per sentence, by applying simultaneously more than one rules to the same sentence, or the same rule more than once, which is permitted by our rule set.

Rule	Frequency
1	0.3
2	0.08
3, 4	0.07
5	0.38
6	0.025
7	0.028
8	0.02
9	0.1

Table 5. Frequency of rule applicability.

Paraphrasing evaluation

A set of experiments have been performed to test the naturalness and the correctness of the final text. Table 5 presents statistical information regarding rule applicability. The frequency column represents the applicability for each rule (the number of times each rule is applicable in the corpus sentences) divided by the sum of the applicability values of all the rules. As can be seen, subject-verb displacement (rule 1) and adverb displacement (rule 5) constitute together around 70% of rule applications.

The 519 instances of paraphrases were checked for grammaticality and naturalness by two native language experts. Table 6 shows the effect of the output sentences on the language experts. The first error rate indicates the percentage of rule applications that have forced the experts to make modifications in order for the paraphrases to become linguistically correct and natural within the initial pool. Modifications entail swaps in the ordering of the chunks. The second error rate is the same percentage for the final pool. Inter-expert agreement exceeded 94%.

Conclusion

The application of shallow paraphrasing rules to Modern Greek sentences for steganographic purposes has been presented. The low paraphrasing level, as well as the absence of any kind of external linguistic resources, enables the easy portability of the methodology to other inflectional languages that are poor in resources. The large average size of paraphrase pools, makes it non-trivial for an unauthorized party to detect the correct paraphrase. An interesting future direction of the current approach would be to take further advantage of the pool size in order to increase the steganographic capacity of the input text.

Rule	Error Rate 1	Error Rate 2
1	10.8%	7.2%
2	41.3%	13.3%
3, 4	21.8%	14.1%
5	15.9%	11.9%
6	0%	0%
7	0%	0%
8	0%	0%
9	3.9%	3.1%
Average	8.2%	5.5%

Table 6. The rules' error rate.

References

Atallah, M., C. McDonough, V. Raskin, and S. Nirenburg. 2000. Natural Language Processing for Information Assurance and Security: An Overview and Implementations. *Proceedings of the Workshop on New Security Paradigms*: 51-65.

Barzilay, R., and L. Lee. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. *Proceedings of the Human Language Technology-NAACL Conference*: 16-23. Edmonton.

Brockett, C., and W. B. Dolan. 2005. Support Vector Machines for Paraphrase Identification and Corpus Construction. *Proceedings of the 3rd International Workshop on Paraphrasing* (IWP). Korea.

Duclaye, F., F. Yvon, and O. Collin. 2003. Learning Paraphrases to Improve a Question-Answering System. In *Proceedings of the 10th Conference of EACL Workshop Natural Language Processing for Question-Answering*. Budapest, Hungary.

Hatzigeorgiu, N., M. Gavrilidou, S. Piperidis, G. Carayannis, A. Papakostopoulou, A. Spiliotopoulou, A. Vacalopoulou, P. Labropoulou, E. Mantzari, H. Papageorgiou, and I. Demiros. 2000. Design and Implementation of the online ILSP Greek Corpus. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*: 1737-1742. Athens.

Kozareva, Z., and A. Montoyo. 2006. Paraphrase Identification on the Basis of Supervised Machine Learning Techniques. *Proceedings of FinTAL, Lecture Notes in Artificial Intelligence*, vol. 4139: 524-533. Springer Verlag, Berlin.

Meral, H. M., E. Sevinc, E. Unkar, B. Sankur, A. S. Ozsoy, and T. Gungor. 2007. Syntactic Tools for Text Watermarking. *Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents IX*. Edited by Delp, Edward J., III; Wong, Ping Wah.

Pang, B., K. Knight, and D. Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. *Proceedings of the Human-Language Technology Conference* (NAACL-HLT). Edmonton, Canada.

Provos, N., and P. Honeyman. 2003. Hide and Seek: An Introduction to Steganography. *IEEE Security and Privacy*: 32-44.

Quirk, C., C. Brockett, and W. B. Dolan. 2004. Monolingual Machine Translation for Paraphrase Generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*:142-149. Barcelona, Spain.

Stamatatos, E., N. Fakotakis and G. Kokkinakis. 2000. A practical chunker for unrestricted text. *Proceedings of the Conference on Natural Language Processing*: 139-150. Patras, Greece.

Topkara, M., C. M. Taskiran, and E. Delp. 2005. Natural Language Watermarking. *Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents*. San Jose.

Toward Automated Stylistic Transformation of Natural Language Text

Foaad Khosmood
University of California Santa Cruz
foaad@ucsc.edu

Robert Levinson
University of California Santa Cruz
levinson@cse.ucsc.edu

Introduction

Style is an integral part of natural language in written, spoken or machine generated forms. Humans have been dealing with style in language since the beginnings of language itself. Almost everyone who is capable of reading and writing, or even just hearing and speaking, routinely identifies and employs different variations of language in daily life. We easily recognize distinct styles of language and can produce our own in multiple variations depending on the context. Computers, however, are not yet capable of anything that sophisticated.

Automatic processing of text styles poses two interrelated challenges: classification and transformation. There have been some recent advances in corpus classification, automatic clustering and authorship attribution of text using a variety of features and techniques [1][9][19][20][21][24][34]. Integral to each approach is a feature set to be extracted (such as n-grams or vocabulary set), and a learning algorithm (such as neural nets or Bayesian methods), to analyze and label the corpora. Project JGAAP [17] has gone one step further and tried to combine and modularize a variety of different classification methods in a standard library.

In contrast to classification; very little research work is available on style transformation. We, can, however study some conceptually adjacent research areas such as natural language generation (NLG) [13][23][26][30], computational linguistics [10], literary studies, stylistics [2][3][6][7][11][28][33], writing assistants [14][15] and machine translation (MT) [8]. In fact, our problem could be likened to statistical machine translation (SMT), except that instead of translating from one language to another, we aim to transform between two styles of the same language. But unlike SMT problems, we do not have the luxury of large pre-existing associative databases such as dictionaries.

Having studies various views and controversies on the concepts of style and stylistics within humanities and linguistics fields, we settle on a definition of style, “as option” and a conscious choice, provided by Walpole [33]. We aim to study natural language styles by building an intelligent, modular and user-friendly system which is capable of making use of a variety of algorithms and methods in order to classify and transform pieces of written text. In this paper we discuss some background work and definitions, and then we present the overall designs for a classification/transformation system. We demonstrate the concept by showing a detectable stylistic shift in a sample piece of text relative to a profiled corpus representing the target style.

Foundational Assumptions

We make a number of simple assumptions which some others have also made in relevant literature. The most basic one is that style can be captured (in a relative sense) using an unbounded set of style markers that can be detected. This assumption allows us to build digital profiles of text around the presence and properties of style markers. A measurable reading of these markers also helps guide style transformation algorithms. We observe that automated rewriting of some parts of the text using some rule-driven rewrite algorithms, can change the statistical style signature of the text and objectively bring it closer to or take further away from a given target corpus signature. To take a trivial example, we imagine a body of text written in modern English and we wish to transform it to Shakespearian (early modern) English. One step among many that would have to be taken would be to transform all modern pronouns to their equivalent early modern ones, such as “you” to “thou”, and “your” to “thine”. This transformation generates a new text which is statistically speaking, closer to the early modern target. Deriving the statistical distance is essentially the same thing as classifying the source text.

Thus, AI system, given text manipulation rules (what we call operators and transforms) and sets of styles to be detected (made up of style makers) should be able to plan a series of operations to manipulate a source text such that it would become as close as practical to a target corpus profile.

Applications

Style classification and transformation have numerous applications in information retrieval (IR), natural language processing (NLP), human computer interaction (HCI), and interactive entertainment. Robust style classification can lead to an entirely new dimension in searching. Not only are new search parameters such as style markers and related tolerance levels possible, but search systems could adopt style searching based on example of an input text. Style comparison techniques could provide for more robust and descriptive plagiarism detection and digital forensics. Individuals, too, can use style transformation software to obfuscate and alter their own style of writing for online privacy reasons. In HCI, richer and more customizable user help interfaces are possible. Texts for such interfaces could gradually adapt to a particular user’s style and facilitate easier understanding. NLG systems such as [13][14][23][26][30] already incorporate some level of style (a high level in case of psychology based [26]) in their generation work, but could enhance their stylistic variability with text-to-text style transformations. In MT [8], automatic style processing could lead to better and more precise translations. Writers and satirists could use style processing to help compose language in deliberate and memorable styles. In interactive entertainment [5], authoring tools for generating game narratives and non-player character (NPC) dialogue could be made much richer, more diverse and more nuanced without significantly increasing a human composer’s burden.

Overview of the Style Recognition and Transformation System

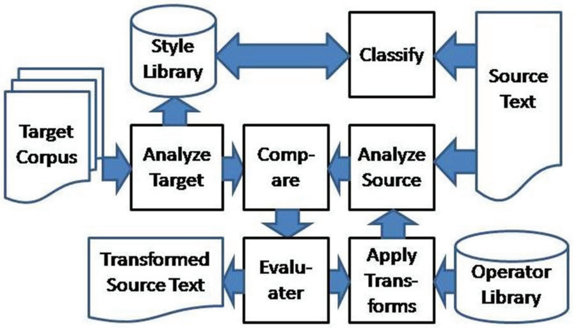


Fig. 1 System Design

The design for our system consists of three major components as shown in Fig. 1. For classification, the system can store marker statistics for all analyzed corpora and it can find closest matches among those for a given source text. For transformation, first the target corpus, typically made of many documents and the source text are both analyzed in terms of the presence of all or a subset of style markers. The system then performs a comparison between the source and target styles, calculates a stylistic distance. The Evaluator decides whether the comparison has yielded a match within pre-defined fuzzy boundaries. If not, the system chooses a transform consisting of one or more operators to apply to the source text in order to do modifications. Once the modifications are done, another comparison is made and if the styles are thought to be still too far apart, more transforms are chosen and applied until the system gets as closest possible to the source style, or if no more transforms are applicable in which case the latest version of the source becomes a "best effort" result of the transformation.

Classification-Transformation Loop

Two databases denoted by cylinders in Fig. 1 symbolize the initial inputs into the system. The ideal situation is if the system already possesses all possible linguistic style markers, and all possible transforms. However, no such comprehensive list exists nor are we likely to construct it. We can begin with a carefully chosen superset of style markers taken from the text classification literature that we have examined but we will surely have to expand it. Similarly, a number of algorithmic text paraphrasing and reformulation rules exist that we can utilize as an initial set of operators. Rather than trying to come up with a huge comprehensive database of markers and transformers, we propose to evolve the list organically as needed during the system development. This is done by a workflow process we call the Classification-Transformation (CT) loop (fig. 2).

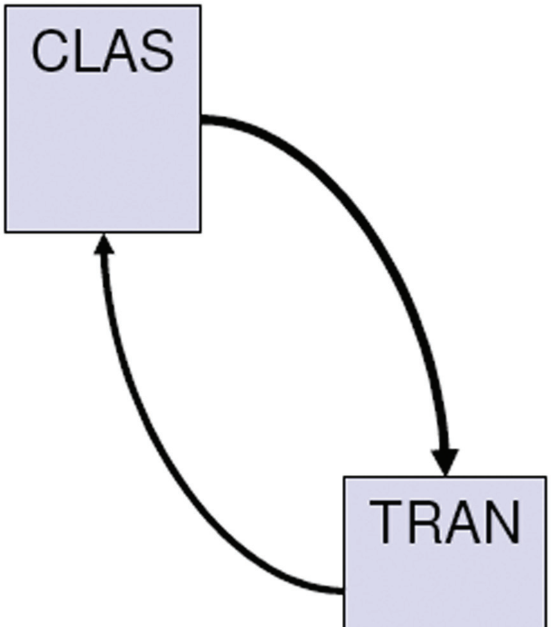


Fig. 2. Classification-Transformation Loop

For example, we might begin with two generic corpus classes: Shakespeare as target and *New York Times* as source. Initially, the classification problem could be very simple, only relying on the presence of the aforementioned early modern pronouns. In the first pass, a *New York Times* source text will be classified as too distinct from the Shakespeare corpus based on the pronoun test. After running a transformation that converts the modern pronouns to their early modern equivalents, the system now will classify the *New York Times* text as Shakespearian. However, we know by human inspection that the source text is far from Shakespearian. This prompts an investigation leading to identification of more distinguishing features of the Shakespearian text that can be extracted and evaluated in the classification process. In essence an expert detected failure is a failure of style specification and it can prompt us to find better markers. This example was trivial because we would never begin with such a limited set of style markers for classification. We envision at least dozens if not hundreds of markers that we can begin with as the initial set. But the problem is essentially the same. Even with hundreds of markers, the system at some point could misclassify a piece of text that is clearly not (yet) of the target style. At this point, it may not be so easy to come up with additional markers yet we know that such markers must exist or else we would not be detecting a misclassification. Going through the CT loop is an organic way to verify the validity of the transformation process and to prompt investigations of ever-more sophisticated markers and transformers. The added advantage here is that each marker and transformer is now available for future transformation exercises. Thus we are strengthening the system as a whole rather than fine-tuning a specific instance of style to style transformation.

Evaluation

A measure of automatic statistical evaluation is already built in to the system at its core. For previous papers, we showed a declining statistical distance curve between detected styles of the source text and target corpora signifying a move in the desired direction. But the nature of this project demands expert human qualitative evaluation at several levels. First, we must verify the correctness and applicability of individual operators. If necessary, we can add additional constraints or evaluations to maximize correctness of the operators. We will have to aim for the broadest possible scope for the operators, while at the same time preserving the grammatical correctness state of the transformed text. Second, post application of operators, we must verify the correctness/coherence of the entire document as whole. Most operators will operate at the sentence level. However, unforeseen cumulative effects are likely. Thus we plan on periodic corpus transformation samples to be evaluated by human experts. We believe for every mismatch, we can add one or more markers to act as discriminators between the misclassified text and the target style. Over time, we can track diminishing levels of corrections conducted by human evaluators.

References

1. Argamon, S., Saric, M., Stein, S., Style (2003) “Mining of electronic messages for multiple authorship discrimination: first results.” Proceedings of the 9th ACM SIGKDD, Washington DC.

2. Biber, Douglas (1989), “A typology of English texts”, Linguistics 27, 3–43.

3. Bradford, Richard (1997), *Stylistics*, part of “The New Critical Idiom” series, Routlidge.

4. Brill, Eric (2000), “Part-of-Speech Tagging”, in Handbook of Natural Language Processing edited by Dale, Moisl and Somer, Marcel Dekker, Inc. 2000, pp 403-414.

5. Bringsjord, S., and Ferrucci (2000), *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, Mahwah, NJ: Lawrence Erlbaum, 2000.

6. Comte de Buffon (1773), “Discourse on Style,” trans. Rollo Walter Brown, in The Writer’s Art, ed. Brown, Harvard University Press, 1921, pp. 285-86. (originally published 1773).

7. Carter, Ronald and Simpson, Paul, Language (1988), Discourse and Literature: An Introductory Reader in Discourse Stylistics, Routledge, 1988.

8. DiMarco, Chrysanne (1994), “Stylistic Choice in Machine Translation,” AMAT.

9. Fakotakis, N. and Stamatatos, E. and Kokkinakis, G. (2001),“Computer-based Attribution without Lexical Measures.” Computers and the Humanities, Volume 35, Issue 2, May 2001, pp. 193-214.

10. Ferrari, Giacomo (2003), “State of the art in Computational Linguistics,” in Linguistics Today: Facing a greater Challenge, International Congress of Linguists, John Benjamins Publishing Company, 2003, p 163.

11. Fish, Stanley (1981), “What is stylistics and why are they saying such terrible things about it”, in Essays in Modern Stylistics, edited by DC Freeman, Routledge, 1981, pp 53-66.

12. Gervas, P. (2000), “Wasp: Evaluation of different strategies for the automatic generation of Spanish verse,” in Proceedings of the AISB00 Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science, 2000.

13. Gon alo Oliveira, Hugo R., Cardoso, F. Amılcar, Pereir, Francisco C., “Tra-la-Lyrics: An approach to generate text based on rhythm,” International Joint Workshop on Computer Creativity, 2007, London.

14. Haardt, Michael (2007), GNU diction(1) PDF manual, accompanying diction version 1.11. 2007. http://www.gnu.org/software/diction/diction.html.

15. Heidorn, George E. (2000), “Intelligent Writing Assistance”, in Handbook of Natural Language Processing edited by Dale, Moisl and Somer, Marcel Dekker, Inc. 2000, pp 181-209.

16. Karlgren, Jussi (2004), “The wheres and whyfores for studying text genre computationally,” In Style and Meaning in Language, Art, Music and Design, Washington D.C., 2004. AAAI Symposium series.

17. Juola, Patrick, et. al. (2009), JGAAP, a Java-based,modular, program for textual analysis, text categorization, and authorship attribution. http://www.mathcs.duq.edu/~fa05ryan/wiki/index.php/Main_Page.

18. Jucker, Andreas H. (1992) *Social Stylistics: syntactic variation in British newspapers*, Walter de Gruiyter, 1992.

19. Kesselj, Vlado et. al. (2003) “N gram-based Author Profiles for Authorship Attribution.” In Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING’03, Dalhousie University, Halifax, Nova Scotia, Canada, August 2003.

20. Khosmood, Foaad and Kurfess, Franz (2005), “Automatic Source Attribution of Text: A Neural Networks Approach,” In IJCNN-05, Montreal, Canada, June 2005.

21. Khosmood, Foaad and Levinson, Robert (2006) “Toward Unification of Source Attribution Processes and Techniques,” IEEE ICMLC, August 2006.

22. Landauer, T. K., Foltz, P. W., & Laham, D. (1998) “Introduction to Latent Semantic Analysis,” 1998, http://lsa.colorado.edu/papers/dp1.LSAintro.pdf.

23. Loehr, Dan., “An Integration of a Pun Generator with a Natural Language Robot,” (1996) In: *Proceedings of the International Workshop on Computational Humor, Enschede*, Netherlands. University of Twente, 1996.

24. Latent Semantic Analysis resources at University of

Colorado, accessed January, 2008. http://lsa.colorado.edu.

25. Luyckx, Kim and Daelemans, Walter (2005), "Shallow text analysis and machine learning for authorship attribution.", In: Computational Linguistics in the Netherlands 2004: selected papers from the Fifteenth CLIN Meeting / van der Wouden Ton [edit.], e.a., Utrecht, LOT, 2005, p. 149-160.

26. Mairesse, Francois and Walker, Marilyn (2008), "A personality-based Framework for Utterance Generation in Dialogue Applications," Proceedings of the AAAI Spring Symposium on Emotion, Personality, and Social Behavior, Palo Alto, March 2008.

27. Moessner, Lilo (2001), "Genre, Text Type, Style, Register, Terminological Maze?" European Journal of English Studies, 2001, Vol. 5, No. 2, pp. 131–138.

28. Murry, John Middleton (1922), *The Problem of Style* (London: Oxford University Press, 1922), p. 77.

29. Rissanen, Matti (1994),"The Helsinki Corpus of English Texts" in Corpora Across the Centuries, Proceedings of the First International Colloquium on English Diachronic Corpora, St. Catharine's College Cambridge, 25–27 March 1993, eds. Merja Kyto, Matti Rissanen and Susan Wright (Amsterdam/Atlanta, GA: Rodopi, 1994), 73–79, pp. 76–7.

30. SciGen -an automatic cs paper generator. 2005, http://pdos.csail.mit.edu.

33. Simpson, John (2004)*, Stylistics: A Resource Book for Students*, Routledge, 2004.

32. Van Sterkenburg, Piet (2003), Editor, *Linguistics Today: Facing a greater Challenge*, International Congress of Linguists, John Benjamins Publishing Company, 2003.

33. Walpole, Jane (1980), "Style as Option," College Composition and Communication, vol. 31, No. 2, Recent Work in Rhetoric: Discourse Theory, Invention, Arrangement, Style, Audience, (May, 1980), pp. 205-212.

34. Whitelaw, Casey and Argamon, Shlomo (2004), "Systemic Functional Features in Stylistic Text Classification", AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design, October 2004.

35. WordNet at Princeton University Cognitive Science Library, http://wordnet.princeton.edu, accessed 9/2008.

Language and Image: T^3 = Text, Tags, and Trust

Judith L. Klavans
University of Maryland, College Park
jklavans@umd.edu

Susan Chun
Independent Museum Consultant, NY
susan@susanchun.com

Jennifer Golbeck
University of Maryland, College Park
golbeck@cs.umd.edu

Dagobert Soergel
University of Maryland, College Park
dsoergel@umd.edu

Robert Stein
Indianapolis Museum of Art
rstein@imamuseum.org

Ed Bachta
Indianapolis Museum of Art
ebachta@imamuseum.org

Rebecca LaPlante
University of Maryland, College Park
laplante@umd.edu

Carolyn Sheffield
University of Maryland, College Park
csheffie@umd.edu

Kate Mayo
University of Maryland, College Park
kmayo1@umd.edu

John Kleint
University of Maryland, College Park
jkleint@umd.edu

In this paper, we present on a new project, "T^3: Text, Tags, and Trust to Improve Image Access for Museums and Libraries", the goal of which is to improve access to digital image collections in museums and libraries for art historians, museum professionals, and the general public. T^3 combines text mining, social tagging, and trust inferencing to enrich metadata and personalize retrieval. We report on an experiment in which users tag a selected set of controversial images; with these tags, similarity profiles are created for subjects to build an initial trust network based on agreement or disagreement. By processing related text through the CLiMB toolkit, we have a third source of evidence for evaluating the role of trust and for assessing the relationship between tags and text terms. We will present collection criteria, including image selection, text identification and choice, and interface choices for data collection and analysis.

The fundamental and driving research issue in this project concerns the relationship between the language of image description and an image itself. The University of Maryland's Institute for Advanced Computer Studies and College of Information Studies, the Indianapolis Museum of Art, and fourteen other museums have joined to conduct research on new methods to improve user access to digital image collections in museums and libraries. Studies on image searching indicate that current subject description and cataloging practices in museums, libraries and other art historical collections are inadequate for many end user needs. Trant, et al. 2007, as part of the steve.museum project, report that search behaviors for users of the Guggenheim collection do not match the descriptive practices of museum personnel. This disconnect results in unsatisfactory and unsuccessful image access for users. Similarly, in Klavans et al. 2008, observations of image cataloging practices in academic visual resource centers reveal that typical records include less than eight subject terms per image with many records containing no subject terms at all. Social tagging is a way to obtain richer, more varied, and more user-oriented subject terms. However, social tagging brings with it the problem of authority and trust: Whose authority is a given user prepared to accept? Whose tags does she want to trust to retrieve relevant images? (Trant and Wyman 2006, Axelrod, Golbeck and Shneiderman 2005).

T^3 is a collaborative, cross-disciplinary project comprised of academic researchers, digital librarians, and museum professionals. We explore the application of techniques from computational linguistics and social tagging to the creation of linkages between the formal academic language of museums and the vernacular language of social tagging. We use text mining algorithms, taxonomies, and lexical resources to identify suggested terms and aid users in tagging images and then retrieving images based on tags assigned from many different perspectives. We use the trust a user places in particular metadata sources, e.g. other users or other sources, to infer a weighted set of results for their searches. Consideration of these weights in ranking algorithms—along with term relationships from lexical resources—has the potential to produce high-quality, focused, personalized retrieval of works from image collections.

The T^3 integrated system builds on three prior research prototypes:

1. CLiMB (**T**ext Mining for Terms): Applies computational linguistic techniques to mine texts associated with images for terms which are then disambiguated, mapped to standard ontologies such as the AAT, and reviewed by museum and library staff for enriching image catalog records with high-quality subject metadata.

2. Steve.museum (**T**agging): Uses on social tagging of images for generating metadata and engaging museum audiences. The current project uses steve tools and methods to explore the roles and usefulness of non-expert enthusiasts in enhancing existing documentation**.**

3. FilmTrust (**T**rust Inferencing): Incorporates trust networks to assign trust values. By gathering input on users' preferred sources, including other users, a trust network automatically assigns values, sourced from both text mining and tagging, based on user perspectives. T^3 will explore the process of extending trust of other users' opinions (i.e., "this user likes the same works I like") to generate values for image descriptions.

This project addresses fundamental research questions in the area of digital image access. Armed with answers to basic research issues, we are able to design environments for improved image access and improved user experience. Research issues include examination of hypothesis to:

- Improve the user experience in finding works of art and interacting with works of art and collections.

- Improve the understanding the relationship between language and visual art, including the use of facets and other knowledge structures to elicit useful tags and assist users in searching

- Examine the relationships, associations, and linkages between terms from different sources, specifically from users, text-mining, and cataloging.

- Study when and how an understanding of sources impacts the value of terms to users and museums, and personalizes the user experience.

Our hypothesis is that through disambiguation and trust information, we can filter out excess terms and rank acceptable terms. This provides users with the capability to adjust their preferred threshold for precision over recall, or the reverse. Specifically, disambiguating terms using a faceted thesaurus provides users with the ability to narrow or expand their searches based on clearly defined concepts. For the trust component, we gather input from users on which sources (people or text) they trust to help us judge how much trust and authority to give to the tags/terms originating from these sources. The trust and authority "ratings" for tags will be used to filter them and/or order the way they are presented. This helps users by showing them the most trusted and authoritative terms first, thus facilitating the user's perusal of query results. Dynamic personalization of these filters helps the user by producing trusted, focused results for queries.

Our initial experiments will explore how users judge trust in this context. Our subjects will tag a series of images and then will rate how much they trust a source (people or text) based on the tags/terms it applied to the same images. Using this data, we will analyze how similarity in tags/terms relates to trust values and if there are particular types of words that have a stronger influence on trust (e.g. emotion words vs. color words). These insights will provide the basis for an initial implementation of our prototype that personalizes search results based on trust. The tagging interface for this experiment is shown in Figure One:

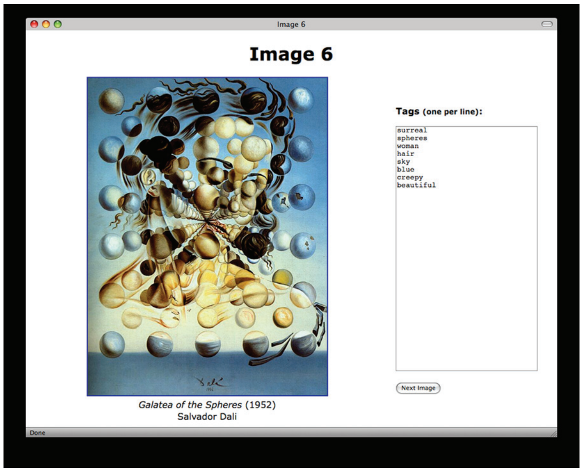


Image 6

Galatea of the Spheres (1952)
Salvador Dali

Tags (one per line):

surreal
spheres
woman
hair
sky
blue
creepy
beautiful

Figure 1. Controversial Image for Tagging for Trust Experiment

T^3 is building an open source working prototype for analyzing and processing terms which serve multiple user communities and allow us to:

- Develop and test new methodologies that group authoritative terms and social tags based on conceptual and semantic relationships
- Test trust-based personalization of results for different user groups
- Research the potential of these new technologies for engaging museum audiences and their impact on the evolving professional landscape of image access

T^3 is funded as a National Leadership Grant by the U.S. Institute for Museum and Library Services. The project is led by Dr. Judith Klavans of the University of Maryland, Robert Stein of the Indianapolis Museum of Art, and Susan Chun, Independent Cultural Heritage Consultant. Dr. Jennifer Golbeck, Assistant Professor in the College of Information Studies at the University of Maryland, is co-PI leads the trust component of the research, and Dr. Dagobert Soergel, Professor in the College of Information Studies at the University of Maryland, leads ontology and knowledge representation aspect of $T^{3.}$ The museum working group is providing users, catalogers, content, and feedback to aid in the research.

Selected References

Axelrod, Adam, Jennifer Golbeck, and Ben Shneiderman (2005), Generating and Querying Semantic Web Environments for Photo Libraries Technical Report, University of Maryland, Department of Computer Science, http://drum.umd.edu.

Trant, Jennifer, David Bearman, and Susan Chun (2007) The eye of the beholder: steve.museum and social tagging of museum collections, in *Proceedings of the International Cultural Heritage Informatics Meeting (ICHIM07)*, J. Trant and D. Bearman (eds). Toronto: Archives & Museum Informatics. 2007. http://www.archimuse.com/ichim07/papers/trant/trant.html.

Klavans, Judith L, Carolyn Sheffield, Eileen Abels, Jimmy Lin, Rebecca Passonneau, Tandeep Sidhu, and Dagobert Soergel (2009) Computational Linguistics for Metadata Building (CLiMB): Using Text Mining for the Automatic Identification, Categorization, and Disambiguation of Subject Terms for Image Metadata. *Journal of Multimedia Tools and Applications*, Special issue on Metadata Mining for Image Understanding (MMIU) 42(1):115-138. Elsevier: Paris.

Trant, Jennifer and Bruce Wyman (2006). Investigating social tagging and folksonomy in art museums with steve.museum. Paper presented at the World Wide Web 2006: Tagging Workshop. http://www.archimuse.com/research/www2006-tagging-steve.pdf.

Mining texts for image terms: the CLiMB project

Judith L. Klavans
University of Maryland, College Park
jklavans@umd.edu

Eileen Abels
Drexel University
eileen.abels@ischool.drexel.edu

Jimmy Lin
University of Maryland, College Park
jimmylin@umd.edu

Rebecca Passonneau
Columbia University
becky@cs.columbia.edu

Carolyn Sheffield
University of Maryland, College Park

Dagobert Soergel
University of Maryland, College Park
dsoergel@umd.edu

The CLiMB (Computational Linguistics for Metadata Building) project addresses the existing gap in subject metadata for images, particularly for the domains of art history, architecture, and landscape architecture. Within each of these domains, image collections are increasingly available online yet subject access points for these images remain minimal. In an observational study with six image catalogers, we found that typically 1 – 8 subject terms are assigned, and that many legacy records lack subject entries altogether. Studies on end users' image searching indicate that this level of subject description is often insufficient. In a study of the image-searching behaviors of faculty and graduate students in American history, Choi and Rasmussen 2003 found that 92% of the 38 participants in their study considered the textual information associated with the images in the Library of Congress' American Memory Collection to be inadequate. The number of subject descriptors assigned to an image in this collection is comparable to what we found in the exploratory CLiMB studies. Furthermore, these searchers submitted more subject-oriented queries than known-artist and title queries. Similar results demonstrating the importance of subject retrieval have been reported in other studies, including Keister, Collins, and Chen 1994.

Under the hypothesis that searchers do not find images they seek partly due to inadequate subject description in metadata fields, the CLiMB project was initiated to address this subject metadata gap by applying automatic and semi-automatic techniques to the identification, extraction, and thesaural linking of subject terms. The CLiMB Toolkit processes text associated with an image through natural language processing (NLP), categorization using machine learning (ML), and disambiguation techniques to identify, filter, and normalize high-quality subject descriptors. Like Pastra et al. 2003 we use NLP techniques and domain-specific ontologies, although our focus is on associated texts such as art historical surveys or curatorial essays rather than captions; unlike generic image search, such as in Google, we analyze beyond keywords and we use text which is specifically and clearly related to an image. For this project, we use the standard Cataloging Cultural Objects (CCO) definition of subject metadata[1] as including terms which provide "an identification, description, or interpretation of what is depicted in and by a work or image."

In order to understand the cataloging process and to inform our system design, we conducted studies on the image cataloging workflow and the process of subject term assignment. Our goal was to collect data on the humanities-driven process as a whole to be able to incorporate our results into an existing workflow and thus assist a portion of the workflow with automatic techniques. An additional purpose of studying the cataloging process was to permit the development of system functionality, i.e., the implementation of rules or the use of statistical methods to identify high-quality subject descriptors in scholarly texts. As part of the CLiMB evaluation, we have established a series of test collections in the fields of art history, architecture, and landscape architecture. These three domains were selected in part because of the existing overlap in domain-specific vocabulary. Testing with distinct but related domains enables us to test for disambiguation issues which arise in the context of specialized vocabularies. For example, the Getty Art & Architecture Thesaurus (AAT) provides many senses of the term *panel* which apply to either the fine arts, architecture, or both, depending on context. In the context of fine arts, *panel* may refer to a *small painting on wood* whereas in the context of architecture, *panel* may refer to a *distinct section of a wall, within a border or frame*.

Figure 1 shows the CLiMB architecture which produces subject term recommendations that can be used into the image cataloging workflow observed in visual resource centers:

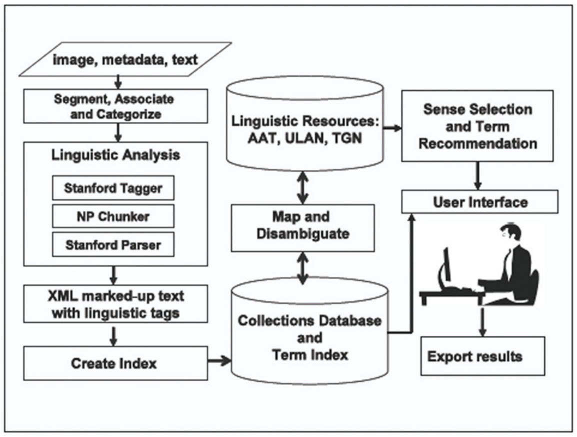

image, metadata, text
Segment, Associate and Categorize
Linguistic Analysis
Stanford Tagger
NP Chunker
Stanford Parser
XML marked-up text with linguistic tags
Create Index
Linguistic Resources: AAT, ULAN, TGN
Map and Disambiguate
Collections Database and Term Index
Sense Selection and Term Recommendation
User Interface
Export results

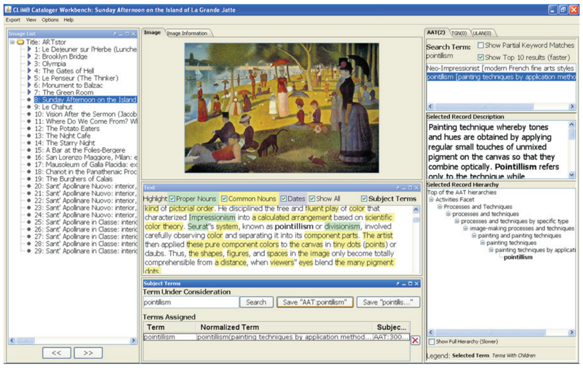
Figure 1. The CLiMB toolkit architecture

CLiMB combines new and pre-existing technologies in a flexible, client-side architecture which has been implemented in a downloadable toolkit and which can be tailored to the user's needs. In addition to matching segments of texts to referenced images, we are developing methods to categorize spans of text (e.g., sentences or paragraphs) as to their semantic function relative to the image. For example, a sentence might describe an artist's life events (e.g. "during his childhood", "while on her trip to Italy", "at the death of his father") or the style of the work ("impressionism"). A set of seven categories – Image Content, Interpretation, Implementation, Historical Context, Biographical Information, Significance, and Comparison – has been initially proposed through textual analysis of art survey texts. These categories have been tested through a series of labeling experiments. Full details are available in Passonneau et al. 2008. The output of this categorization will be incorporated in future versions of the Toolkit, and will be used as part of the disambiguation component.

An important contribution of the CLiMB project is the development of a disambiguation component, enabling the system to move beyond standard keyword-based indexing by associating words and terms that have multiple meanings which correspond to different descriptors with the correct meaning in context. The ability to select one sense from many is referred to as lexical disambiguation. Results of our ongoing studies on sense disambiguation using hierarchically structured faceted thesauri and lexical resources, such as the Art and Architecture Thesaurus and WordNet, will be presented**.** We have experimented with the use of WordNet, with different levels of the facets of the AAT, and with different degrees of filtering for modifiers in noun phrases. We also have results on setting weights for each of these factors to determine the most accurate disambiguation techniques.

One of the most vexing problems in word sense disambiguation is the fact that often several senses could be considered correct within a given context. Therefore, evaluation can be a challenge since there may be no clear-cut right or wrong. The need for fuzzy evaluation will be discussed in our presentation, with a demonstration of different ways to measure precision and recall against a "moving target" baseline.

Figure 2 shows the CLiMB interface in its current state as of Fall 2008; as we use the results of our experimental research, this interface may change as of the time of presentation of the paper.


CLiMB Cataloger Workbench: Sunday Afternoon on the Island of La Grande Jatte
Export
View
Options
Help
Image List
Image
Image Information
Text
Highlight
☑ Proper Nouns
☑ Common Nouns
☑ Dates
☑ Show All
☑ Subject Terms
Subject Terms
Search
Save "AAT:pointilism"
Save "pointilis..."
<<
>>
AAT(2)
TGN(0)
ULAN(0)
□ Show Partial Keyword Matches
☑ Show Top 10 results (faster)
Selected Record Description
Selected Record Hierarchy
□ Show Full Hierarchy (Slower)
Legend: Selected Term Terms With Children

Title: ARTstor
1: Le Dejeuner sur l'Herbe (Lunche
2: Brooklyn Bridge
3: Olympia
4: The Gates of Hell
5: Le Penseur (The Thinker)
6: Monument to Balzac
7: The Green Room
8: Sunday Afternoon on the Island
9: Le Chahut
10: Vision After the Sermon (Jacob
11: Where Do We Come From? Wl
12: The Potato Eaters
13: The Night Cafe
14: The Starry Night
15: A Bar at the Folies-Bergere
16: San Lorenzo Maggiore, Milan: e
17: Mausoleum of Galla Placidia: ex
18: Chariot in the Panathenaic Proc
19: The Burghers of Calais
20: Sant' Apollnare Nuovo: interior,
21: Sant' Apollnare Nuovo: interior,
22: Sant' Apollnare Nuovo: interior,
23: Sant' Apollnare Nuovo: interior,
24: Sant' Apollnare Nuovo: interior,
25: Sant' Apollnare in Classe: interic
26: Sant' Apollnare in Classe: interic
27: Sant' Apollnare in Classe: interic
28: Sant' Apollnare in Classe: interic
29: Sant' Apollnare in Classe: interic

kind of pictorial order. He disciplined the free and fluent play of color that characterized Impressionism into a calculated arrangement based on scientific color theory. Seurat"s system, known as **pointillism** or divisionism, involved carefully observing color and separating it into its component parts. The artist then applied these pure component colors to the canvas in tiny dots (points) or daubs. Thus, the shapes, figures, and spaces in the image only become totally comprehensible from a distance, when viewers" eyes blend the many pigment dots.

Term Under Consideration
pointilism

Terms Assigned

Term	Normalized Term	Subjec...
pointillism	pointillism(painting techniques by application method...	AAT:300...

Search Term:
pointilism

Neo-Impressionist [modern French fine arts styles
pointilism [painting techniques by application metho

Painting technique whereby tones and hues are obtained by applying regular small touches of unmixed pigment on the canvas so that they combine optically. **Pointillism** refers only to the technique while

Top of the AAT hierarchies
- Activities Facet
  - Processes and Techniques
    - processes and techniques
      - processes and techniques by specific type
        - image-making processes and techniques
          - painting and painting techniques
            - painting techniques
              - painting techniques by applicati
                - **pointillism**

Figure 2: The CLiMB Toolkit Interface

Note in Figure Two that the collection under review is found in the left panel, the image is in the center, the analyzed text is shown to the cataloger, with the searchable Getty thesaural resources (AAT, Thesaurus of Geographical Names (TGN) and Union List of Artist Names (ULAN)) in the right panel. The cataloger can select subject terms, and when possible, normalize according to the Getty unique identifier. All interface panes are flexible, and can be hidden or enlarged, as required by the user. Cataloger subject term selections can be exported in a range of formats (see Export button in upper left hand corner of Figure 2) for incorporation into an existing catalog record.

To sum, in this paper we will present:

- The problem of subject term access in image retrieval
- **The CLiMB system**, which utilizes computational linguistics and machine learning to improve basic keyword search through:
  - Semantic categorization of text segments
  - Disambiguation

- User evaluation studies and findings

Selected References

Chen, H. (2001) An Analysis of Image Retrieval Tasks in the Field of Art History. Information Processing & Management, Vol. 37: 701-720.

Choi, Y. and E. Rasmussen (2003) Searching for Images: The Analysis of Users' Queries for Image Retrieval in American History. Journal of the American Society for Information Science and Technology, Vol. 54: 498-511.

Collins, K. (1998) Providing Subject Access to Images: A Study of User Queries. The American Archivist, Vol. 61: 36-55.

Keister, L.H. (1994) User Types and Queries: Impact on Image Access Systems. In: Fidel, R., T.B. Hahn, E. Rasmussen, P. J. Smith (eds.): Challenges in Indexing Electronic Text and Images. Learned Information for the American Society of Information Science, Medford: 7-22.

Klavans, Judith L, Carolyn Sheffield, Eileen Abels, Jimmy Lin, Rebecca Passonneau, Tandeep Sidhu, and Dagobert Soergel (2009) Computational Linguistics for Metadata Building (CLiMB): Using Text Mining for the Automatic Identification, Categorization, and Disambiguation of Subject Terms for Image Metadata. Journal of Multimedia Tools and Applications, Special issue on Metadata Mining for Image Understanding (MMIU) 42(1):115-138. Elsevier: Paris.

Passonneau, R., T. Yano, T. Lippincott, J. Klavans (2008) Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning. 3rd International Conference on Computer Vision Theory and Applications, Workshop on Metadata Mining for Image Understanding: 13-22.

Pastra, K., H. Saggion, Y. Wilks, (2003) Intelligent Indexing of Crime-Scene Photographs. In: IEEE Intelligent Systems: Special Issue on Advances in Natural Language and Processing, Vol. 18, Iss. 1: 55-61.

Notes

[1] http://vraweb.org/ccoweb/cco/parttwo_chapter6.html.

Library Collaboration with Large Digital Humanities Projects

William A. Kretzschmar, Jr.
University of Georgia
kretzsch@uga.edu

William G. Potter
University of Georgia

At DH2007 a special session on "finishing" large humanities research projects (now forthcoming as a cluster in *DHQ*) suggested, in part, that particular stages of such projects might be completed, but that continuing institutional support was important for the long-term sustainability of the projects and their products. At DH2008 a special session was devoted to "Aspects of Sustainability in Digital Humanities," in which technical, organizational, and scholarly dimensions were discussed with reference to a museum project, along with metadata and the issue of portability in other settings. In this paper, we would like to continue the theme of sustainability. We will discuss issues of institutional support for a large digital humanities project, and then propose collaboration with the university library as the only realistic option for long-term sustainability in our environment. We believe that our experience is typical of the situation for other projects, large and small, that many digital humanities faculty now face at their institutions, and therefore that our experience is also typical of the demands that will be placed on libraries to sustain faculty digital research for the long-term.

As for many digital projects, the Linguistic Atlas Project (LAP) began with computing resources located within the research office itself, first personal computers and later servers. When the university created a research computing service (as an addition to the institution's administrative and instructional services), LAP was one of the first clients--the editor of LAP was even asked to help design the service. However, over the course of several years the funding structure for the research computing service changed from an essentially institutional budget with additions from externally funded research, to a fee-based service much more dependent on research with annual external funding. This meant that humanities projects like LAP, while not excluded from the research computing service, either needed to find consistent external funding or hope for sufferance from the paying customers. Neither of these options appeared likely to sustain LAP digital products and archives over the long term, and, citing budget pressures, the university administration declined to guarantee funding for gaps in the external funding that LAP has regularly solicited and received. LAP reacquired a server for its research office and 20 Tb of storage space for its large and growing multimedia archive.

Enter the University Librarian. During discussions about institutional support for LAP, whose paper and audio tape records are stored in a library special collections facility, the University Librarian suggested that the library was expanding its multimedia collections (e.g., our institution hosts the Peabody awards, and so has an important interest in archiving TV materials), and even the large multimedia collections of LAP could be incorporated there. Unlike the research computing service, the mission of the library includes archiving and dissemination, now increasingly of digital materials as well as traditional paper. We are now developing cooperation not just for a multimedia archive, but also for dissemination of LAP products and information through library means in the context of continuing scholarly activity.

The terms of our cooperation can be discussed under just the headings of the DH2008 sustainability session:

—Technology. Two issues dominate here, long-term storage and security. It had been planned that LAP materials would be stored on spinning disk in the research computing service storage array. However, the library holds its multimedia collections on LTO-4 tape. The latter is much more cost effective; however, tape storage requires periodic refresh of the medium (which the library will carry out indefinitely as part of its archival mission), and limits the speed of access to the materials. LAP and the library intend to disseminate the project's multimedia (audio, and later image) products, but the speed of online access is not required for all materials. We will discuss what can be disseminated online and by other means and the terms of such dissemination, as well as the balance between spinning disk and tape in a setting like ours.

Security is an issue for the online presence of LAP at the library. The project has had a heavily interactive Web presence including GIS for over a decade, first on servers in the research office and later at the research computing service. However, the library has not provided interactive Web access of this kind before. We will discuss integration of LAP programming into the library server system, including reversal of the previous practice of heavy server-side scripting to place more emphasis on client-side processing.

—Organization. LAP and the library will use metadata standards developed for library and linguistic collections, but the question of standards is not trivial because metadata practices for libraries and linguists have not developed in parallel owing to different goals of the practices. We will discuss the resolution of these differences.

We will also discuss the organization of the partnership of the library and the LAP. Just as for the research computing service, there are real costs associated with continuing services provided to LAP by the library. However, the difference in the mission of the research computing service as compared to the library makes a difference in how such costs might be borne over the long term. Periodic infusions of external funding can assist the library to accomplish its goals, but the improved match with LAP as an archival and information resource makes it more reasonable to provide continuing support for the project without annual external funding.

—Scholarship. LAP cannot just be given over to the library in its entirety because it remains very active in scholarly work. As long as LAP continues to conduct new research, it can cooperate with the library to keep its Web presence up to date including both links and programming for interactivity. The LAP archives can be extended with new material, but also with new versions of old materials (such as digital images of paper records). We will discuss how new work and additional aspects of older materials can continue to round out LAP collections in the library, as an active partnership instead of a static archive.

We believe that cooperation at our institution between the library and a large digital humanities project should not be a singular occurrence. Some digital humanities projects have always been associated with and supported by the university library, and many libraries, including ours, have been active with their own digital initiatives. However, a great many more digital humanities projects began as separate faculty initiatives, and many libraries are now interested in developing their digital archives and activities. If independent projects are to be sustained beyond their initial development, they will require the sort of new partnership that we discuss. Libraries, too, can enhance their digital activities through cooperation with faculty digital humanities projects as information resources. We hope our partnership can provide a model.

Babylon: Displacement and Re-creation of Calderón's *Life is a Dream*

Elizabeth Sofia Lagresa
University of California, Santa Barbara
elagresa@gmail.com

With the emergence of digital texts and the movement to digitize existing texts, a complementary push to develop digital means of textual analysis continues to surface. New digital analysis tools can draw on large databases of texts or apply multiple analysis methods to a single text in minutes. Thus, digital textual analysis potentially provides the opportunity to examine large amounts of texts in scant time (Moretti, 2007), while offering visualization potentials that exceed that of texts. For example, these tools can create graphs of word occurrences, map those occurrences according to concordance, or remove all occurrences of a word to show its frequency in a text.

Described by Rockwell (2003) as exploring "the question of the relationship between how we represent texts, how we see them, and our theories of textuality," textual analysis generally seeks to identify patterns within the text, such as concordance or unity (Rockwell, 2003), meaning (Samuels & McGann, 1999), truth (Brooks, 1947), or rhetorical strategy (Bazerman & Prior, 2004). Yet, another implication of digital textual analysis is that it involves the re-creation of a text. This provides an opportunity for us to develop a heightened understanding of what is the role of readers and writers as co-creators of literature, and establishes the theoretical basis necessary to link textual analysis with translation. As its Latin root *translatio* (to transfer, to carry, to bring across, to displace) suggests, translation's basic function is to move meanings from one context to another; consequently, translation can denote concepts such as paraphrase, decoding, interpretation, communication, and even re-creation.

Borges' theory of translation, as illustrated in his famous works "Pierre Menard, Author of the Quixote" and "The Translators of the One Thousand and One Nights," opens up a world of limitless possibilities by elevating translation from an act of reproduction to one of re-creation. In the fictional short story *Pierre Menard,* Borges highlights the importance of the context, the reader and the translator as co-creators of meaning, ushering a new theory of literary criticism and translation that trivializes the preoccupation with notions such as faithfulness, authorship and originality. He supports this theory in his essay on the translators of the *Arabian Nights*, by making the translator's infidelities into what is most important and valued. As a result, he demonstrates that literary translations can produce diverse representations of the foreign text and culture, which ultimately enrich both languages and texts through the act of remaking.

Utilizing Borges' theory as the conceptual backbone, this project blends traditional and digital methods of textual analysis to create visualizations (tag clouds, word trees, influential word maps) that serve as entry points for analysis. With the goal in mind of investigating translation theory, the paper studies the effects of man-made, machine-made translations, and digital text-analyses of a passage from Spanish Golden Age drama. This approach provides a greater understanding of how Spanish source editions (Cruickshank, and Williamsen), man-made translations (Racz, FitzGerald, and MacCarthy), and machine-made translations (Babylon software) versions of a soliloquy from Calderón's play "La vida es sueño," compare and contrast to one another. To re-create both source and man-made translation texts the Babylon translation software tool was applied, first to convert them from Spanish into English, or English into Spanish, and then to translate those versions back into the original language, Spanish or English. After completing the first level operation of running the source text repeatedly and recursively through the Babylon translation software application, which utilizes a statistical approach to automatic language translation and natural language processing, the second level operation was set in motion. This level involved decomposing all online texts (source editions, man-made translations, machine-made translations) into their constituent elements and patterns with the aid of text-analysis and statistical visualization tools. In order to compare and contrast all versions amongst themselves and with each other, various text-analysis tools, namely TagCrowd, TAPoR, ManyEyes and Crawdad were utilized to visualize word patterns, themes, and word frequencies along with their importance. It is significant to note that the tools trended toward certain forms of interpretation, and that these trends affected the types of analysis possible, but, more importantly, they could also be re-purposed for alternative approaches.

The two tools that were found to be the most useful for the comparison of different versions, as well as for their application to Spanish texts, were ManyEyes and Crawdad. ManyEyes was employed to generate tag clouds and word trees. While tag clouds were used to visualize word frequencies with number of occurrences in their context, which is useful in comparing word usage patterns; word trees were particularly helpful in visualizing individual words, phrases and punctuation concordances show within their respective context in order to reveal recurrent themes. Crawdad, on the other hand, was employed to visualize a network of the most influential words and their interconnections, along with their frequency of co-occurrence. The tool also provided influence scores, which served as indicators of high coherence and focus in the text.

The overall intent of the project is not to determine which version is more faithful to the "original," but rather to explore how these creative "infidelities" and re-creations enrich both languages and cultures. Through translation and digital textual analysis we can illuminate the challenges and possibilities presented by the transformation/ displacement of seemingly discrete national, cultural and literary territorialities. For example, of particular interest in the research results is the appearance of deviations in minor themes, which demonstrate that, although, main themes predictably correlate closely across versions, subordinate themes follow less predictable patterns. By studying these variations we can explore how each re-creation reflects the intersection between languages, as well as the existence of latent possibilities of diverse meaning contained within the source text. The result is the development of a networked fidelity framework, which rejects the traditional faithful/unfaithful binary that has predominated in translation studies.

The project corroborates the impossibility of a word for word translation and even edition, and further supports the acceptance of multiple versions as all simultaneously genuine and divergent, while highlighting the difficulty and futility of placing versions in a hierarchy. Additionally, since all translations must be equally encompassed as "unfaithful," they can also be judged as "faithful," highlighting the fact that translation is an act of subjective interpretation and re-creation. Consequently, translation (be it machine or man-made) shares similarities not only with text-analysis, but also with deformance, and the profound subjectivity from which critical insight emerges. This posits that translations not only have the power to share the work with a wider culture, but also to enrich both languages and texts through the act of remaking. This, in turn, establishes the possibility of equating translation to a creative process that results in unique text, eliminating the hierarchical divide that exists between author and translator, and original and translation, as envisioned by Borges' theory of translation.

Furthermore, the project shows that digital visualization tools for "distant" reading and traditional methods of close reading do inform and complement each other to expand our understanding of all texts, including those in translation. However, do to their dependence on each other they do require flexibility in order to switch back and forth between both modes of interpretation. Lastly, there are many possible ramifications of this research project. For example, the use of machine-assisted translation for the *analysis* (and *interpretation*) of translations, not just for the actual generation of translated texts, as well as an alternative way to explore how meaning is constructed and re-constructed in literature.

Works Cited

Babylon: Translation at a single click. 1997-2007. Babylon Ltd. 13 Feb. 2008. <http://www.babylon.com/>.

Bazerman, Charles and P. Prior *What Writing Does and How it Does it*. Mahwah, New Jersey: Lawrence

Earlbaum Associates, 2004.

Borges, Jorge Luis. "Pierre Menard, Author of the *Quixote*." *Ficciones*. English Trans. Buenos Aires: Editorial Sur, 1944. 88-95.

- - - . "The Translators of the One Thousand and One Nights." Trans. Esther Allen. *The Translation Studies Reader.* Ed. Lawrence Venuti. 2nd ed. New York: Routledge, 2004. 94-108.

Brooks, Cleanth *The Well Wrought Urn: Studies in the Structure of Poetry*. Orlando, Florida: Harcourt Brace & Company, 1947.

McCarty, Willard, Humanities Computing New York: Palgrave MacMillan, 2005.

Moretti, Franco. *Graphs Maps Trees: Abstract Models for a Literary History*. London: Verso, 2007.

Samuels, Lisa Jerome McGann. "Deformance and Interpretation." *New Literary History,* 30.1 (1999): 25-56.

Rockwell, G. "What is Text Analysis, Really?" *Literary and Linguistic Computing,* 18 (2003): 209-219.

A framework for multilayered boundary detection: initial results from the Clementine Vulgate

Thomas Lippincott
Columbia University
tom@cs.columbia.edu

Introduction

We present a framework for general boundary detection in texts with complex compositional histories. The framework is designed for end-to-end testing of hypotheses via linguistic feature extraction and machine learning. We describe initial results on the Vulgate Bible utilizing the inflectional richness of the Latin language and several well-known facets of its composition. These results indicate that the framework is an effective testbed for theories in source criticism, and we propose further work that would extend its functionality to more texts and facets.

Texts with a history as rich as the Bible present a unique opportunity to study the interaction of compositional features. Scholarship ranges from consensus on fundamental points, to competing theories in source criticism and translation. Moreover, passages in the Bible have been grouped by style (poetic, historical, legal), function (apocalyptic, prophetic), traditional author (Moses, Joshua) historical time period (Torah, Lamentations) and so forth. It is less clear what, if any, practical linguistic differences these groupings represent, and how they have interacted over time. We consider several widely-accepted scholarly beliefs in choosing the targets for our preliminary machine-learning experiments.

Data

We perform proof-of-concept experiments on the Clementine Vulgate, the official canon of the Catholic Church from 1592 to 1979, because of the relative uniformity and well-documented history of the text. The Vulgate is composed entirely in Latin, a highly-inflectional liturgical language of the Catholic Church and medieval scholarship. Its regular, rich morphology makes it very amenable to computational linguistics, although as a liturgical language it receives little attention in practical contexts. Every non-function word is distinguished by a suffix which indicates grammatical qualities like gender, number, tense, voice, etc. as well as syntactic role. Most words belong to one of a small number of classes for which these endings are completely deterministic: for example, nouns belong to one of five declensions, while verbs belong to one of four conjugations. Strict agreement between parts of speech makes word-order almost irrelevant, semantically. The text itself was composed circa 400 A.D. by Jerome, from Greek, Hebrew, Latin and Aramaic sources, and is accompanied by his commentary on his translation methodology.

Non-traditional literary studies

Before presenting the framework, we address some common pitfalls that arise when applying computational methods to an ancient text, and how we attempt to avoid them. Rudman[13] gives an overview of inherent problems in such studies: of these,we are particularly concerned here with addressing the following: knowledge of the disciplines that make up the field and incomplete selection of style markers.

To avoid the errors of the interloper, we keep language- and domain-specific choices distinct from our general framework. Our principles are simply a) the text of the Bible can in principle be divided along many historical dimensions, b) linguistic features may remain that indicates these divisions, c) machine learning, based on these features, will be more successful at learning valid than invalid divisions. These, we feel, are unbiased general assumptions that lay the groundwork for collaboration with domain experts.

The pitfalls in feature selection ("style markers") include limited feature sets and unfounded generalisations about feature relevance (i.e. "style as a monolithic concept"). We are very conscious of this, and in fact an initial motivation for the study was to investigate the heterogeneous usage of "style" in a text that demonstrates so many. We throw a wide net in feature extraction, and present our reasoning for subsequent modifications to this set.

Finally, Rudman[13] argues that non-traditional (i.e. computational) studies should only follow extensive traditional studies. This criterion is certainly met here: in fact, our results so far are entirely based and evaluated upon hypotheses developed over the past two centuries of Biblical criticism, and concludes with an in-depth application to a dominant theory in the field.

Framework

We will present our framework in detail: the major points are that it is written in the Python programming language, uses TEI-derived document encodings, and uses the WEKA toolkit for machine learning. Primary concerns are generality and modularity: specifically, the feature extraction methods are simple APIs that can easily be extended across languages. We use a simple notation to create “hypothesis-files” that capture a single testable theory for a classifier to attempt to learn. This notation is capable of fine-grained divisions, down to the level of individual words. It can also incorporate multiple primary sources and languages.

Features

The simplest feature set we consider are lemma frequencies. These perform superbly, as they train against topical lemmas. To prevent arbitrary over-training, such as using proper names as features, we only retain words recognized by a general-purpose Latin dictionary. Still, overtraining to the narrative remains a concern, and for that reason we do not focus on this feature set. A reasonable approach may be to only use non-topical function words, which were found to perform well in a study by Garcia[4].

The second feature set is part-of-speech frequencies. Latin words all begin as noun, verb or adjective, and through inflection take on diverse parts of speech. Since it is somewhat arbitrary how we distinguish parts of speech and inflected forms here, we have begun with the extremes of the fundamental types and the fully-inflected forms.

The third feature set is inflectional frequencies. Collatinus [10] is capable of lemmatising latin words and preserving the inflectional information that is stripped off. We find 634 different inflection types throughout the Vulgate. Unlike in most languages, an isolated word in Latin can show an unambigious syntactic role via inflection. Therefore, this feature set includes syntactic labels, which proved useful for Hirst et al[6] when extracted as bigrams. We have not attempted this yet, as the inflectional analysis needs to be improved first. This is an important issue that we will discuss in-depth.

Applications

For the proof-of-concept, our targets for machine learning are relatively undisputed divisions of the text. For example, we consider language (immediately prior to the Vulgate), literary style, and original author. Dividing the text according to these features, the boundaries usually fall between books. There are exceptions to this, for example a passage in Esther known to have been written separately, but for the purposes of our initial experiments we divided the text into sets of books. The results confirm the framework’s ability to detect stylistic boundaries, and careful examination of its “misclassifications” sometimes reveal subtle textual affinities. Friedman[3] has presented a fine-grained theory of the composition of the Torah, and we intend to encode this in our hypothesis. Space and time permitting, we will present the results of several variations of the theory as applied to the Hebrew Masoretic text of the Torah.

References

[1] Gregory R. Crane. Perseus, 2008. http://www.perseus.tufts.edu/.

[2] William G. Dever. *What did the biblical writers know, and when did they know it?* Eerdmans Pub., Cambridge, UK, 2001.

[3] Richard Elliott Friedman. *The Bible with Sources Revealed*. HarperCollins, New York, NY, 2003.

[4] Antonio Miranda Garcia and Javier Calle Martin. Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1), 2007.

[5] Neil Graham, Graeme Hirst, and Bhaskara Marthi. Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4):397–415, 2005.

[6] Graeme Hirst and Ol’ga Feiguina. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417, 2007.

[7] Reinhard G. Kratz and John Bowden. *The Composition of the Narrative Books of the Old Testament*. Vandenhoeck and Ruprecht, Göttingen, Germany, 2005.

[8] Edward Loper and Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, pages 63–70, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[9] Bruce M. Metzger and Roland E. Murphy.
The New Oxford Annotated Bible. Oxford University Press, New York, NY, 1994.

[10] Yves Ouvrard. Collatinus, 2005. http://www.collatinus.org/.

[11] W.E. Plater and H.J. White. *A grammar of the Vulgate, being an introduction to the study of the latinity of the Vulgate Bible*. Oxford at Clarendon Press, 2nd edition edition, 1926.

[12] J. Platt. Machines using sequential minimal optimi-

zation. In B. Schoelkopf,

C. Burges, and A. Smola, editors, *Advances in Kernel Methods -Support Vector Learning*. MIT Press, 1998.

[13] Joseph Rudman. Non-traditional authorship attribution studies in the historia augusta: Some caveats. *Literary and Linguistic Computing*, 13(3), 1998.

[14] M. Seutter, O. Seibert, and C.H.A. Koster. Agfl: Affix grammars over a finite lattice, 2005. http://www.agfl.cs.ru.nl/.

[15] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition edition, 2005.

Chasing the Ghosts of Ibsen: A computational stylistic analysis of drama in translation

Gerard Lynch
Trinity College, Dublin
gplynch@tcd.ie

Carl Vogel
Trinity College, Dublin
vogel@tcd.ie

1 Introduction

Research into the stylistic properties of translations is an issue which has received some attention in computational stylistics. Previous work by Rybicki (2006) on the distinguishing of character idiolects in the work of Polish author Henryk Sienkiewicz and two corresponding English translations using Burrow's Delta method concluded that idiolectal differences could be observed in the source texts and this variation was preserved to a large degree in both translations. This study also found that the two translations were also highly distinguishable from one another.

Burrows (2002) examined English translations of Juvenal also using the Delta method, results of this work suggest that some translators are more adept at concealing their own style when translating the works of another author whereas other authors tend to imprint their own style to a greater extent on the work they translate.

Our work examines the writing of a single author, Norwegian playwright Henrik Ibsen, and these writings translated into both German and English from Norwegian, in an attempt to investigate the preservation of characterization, defined here as the distinctiveness of textual contributions of characters.

2 Background

Many studies in computational stylistics have focused on tasks which are related to those of authorship attribution but are not concerned with the notion of attributing authorship to texts of unknown provenance. A related area of study is the idea of *pastiche*, an intended imitation of an author's style in the same language, which contrasts with translation as an intended imitation of an authors style but in a different language. Somers and Tweedie (2003) conducted experiments involving pastiche, the author in question was Lewis Carroll and the pastiche was a modern children's fable written by Gilbert Adair called *Alice through the Needle's Eye* in which the author attempted to imitate the style of Carroll in such works as *Through the Looking Glass* and *Alice's Adventures in Wonderland*. Various techniques used in authorship attribution were used in the task, including methods of lexical richness, principal component analysis, the cusum technique[1], and others. Some methods distinguished the pastiche from the original and some did not. Somers and Tweedie (2003) conclude as follows: If a pastiche is indistinguishable from the original by an authorship attribution method, can it be said that the pastiche is in fact a perfect imitation of the original, or is it the case flawed? In the case of translation which is of relevance to our current work, the question can be formulated in a different way: If a translation is highly similar stylistically to other works by the same translator, is the translation a faithful one?

This current study builds on previous work detecting character voices in the poetry of Irish poet Brendan Kennelly by Vogel and Brisset (2007) and a study on characterization in playwrights by Vogel and Lynch (2008). These studies were concerned with the language used by authors in the creation of character. The tools used in this study were used in these previous studies.

3 Experimental Setup

For these experiments, three works by Henrik Ibsen were used, *A Doll's House* (1879) *Ghosts* (1881), and The Master Builder (1892)[2] . The electronic versions of these plays were obtained from Ibsen.net[3] and Project Gutenberg. The contributions of each character are extracted using PlayParser[4] . All stage instructions are discarded in this step, leaving only the remaining character dialogue. The method decomposes all texts associated with a category (here, persona or play) into chunks of equal size. Pairwise similarity metrics are computed for all chunks. The metric is just the average chi-square computation of the difference in distribution between pairs of files for each token appearing in either file. Different sorts of tokenization capture different linguistic features for which one might consider distributions within and across text categories. If the pairwise similarity scores are rank ordered, then one can exploit the intuitions that a homogeneous category will have a smaller rank-sum than a heterogeneous one, and that arbitrary samples from a homogeneous category should be more like the rest of that category than alternative categories. The method also provides a way to measure degree of homogeneity, the number of samples who are more like the rest of the category can be measured against a baseline creating by random sampling. See Vogel and Lynch (2008) for a more detailed account of the method.

4 Experiments

4.1 First Experiment

The first experiment seeks to compare character homogeneity over different languages. The second experiment compares two different translations of the same play in order to quantify similarity between parallel translations. Table 1 shows the plays and their respective translators. As mentioned, the first 10k of text per character was examined and this was split into 5 sections. Thus, the criteria for inclusion in the study was that the character should contain at least 10k of text and 11 characters were examined, as detailed in Table 2. Only the version of Ghosts translated by Archer is used in the first experiment. The results named in the next section have statistical significance.

Play	**Language**	**Translator**
Gespenster (*Ghosts*)	German	Sigurd Ibsen
Ein Puppenhaus (*A Doll's House*)	German	Marie Von Borch
Baumeister Solness (*The Master Builder*)	German	Marie Von Borch
The Master Builder	English	William Archer & Edmund Gosse
A Doll's House	English	William Archer
Ghosts	English	William Archer
Ghosts	English	R Farquarson Sharp

Table 1: Plays and Translators

The results for the first experiment showed that character homogeneity varies to some extent over the translations, the character idiolects are not necessarily preserved to the same degree as the originals. When letter frequencies are measured, the Norwegian original language characters prove to be more homogeneous than the translations, examples include the character of Engstrand who is homogeneous in English and Norwegian but not German, however, one character whose language remains distinct across all of the translations is Nora, the heroine from *A Doll's House* and one of the typical strong female characters found in Ibsen's drama.[5] However, when the play is taken as the category, we find that the chunks of personas from each play are more similar to the personas from the same play than from different plays, and this is consistent across languages. So while within character homogeneity is not always preserved, the homogeneity of the plays remains relatively consistent across languages.

5 The Second Experiment

The second experiment sought to examine whether two translations of the same original text into the same language are distinguishable by translator as in the work by Rybicki which delineated the work by each, while observing the preservation of idiolect in each. The experimental setup was similar to the first experiment with the character contributions separated and split into five files each. This time, however, the characters from the two translations of *Ghosts* by William Archer and Robert Farquharson Sharp were compared with each other.

Our findings were that the characters from Archer's translation were more homogeneous in general than those of Sharp's translation. Of the characters which were not homogeneous, the text segments were more similar to the segments of the same character by the corresponding author than any other writings by the same author. Sharp's characters tended to be more similar to the corresponding Archer character more often than vice versa. This suggests that both authors have managed to perform faithful translations which are not highly influenced by their own writing style. It also suggests that Sharp may have used Ibsen's translation as a reference when crafting his own.

Character	Play
Engstrand	*Ghosts*
Pastor Manders	*Ghosts*
Oswald	*Ghosts*
Mrs Alving	*Ghosts*
Helmer	*A Dolls House*
Krogstad	*A Dolls House*
MrsLinde	*A Dolls House*
Nora	*A Dolls House*
Aline	*The Master Builder*
Hilde	*The Master Builder*
Solness	*The Master Builder*

Table 2: Characters and their plays

This result contrasts with Rybicki (2006) who found that the two translations of Sienkiewicz separated cleanly from one another with a preservation of individual character idiolects. However, Rybicki makes clear that the two English translations were done almost one hundred years apart with the second translator taking specific steps to bring the language of Sienkiewicz into the 20th century. Also, we are aware that results between the studies of two different authors are not directly comparable and do not seek to draw definite paral[6]lels, merely to reflect on related work in the same sphere.

6 Conclusion

In this research, character idiolects in translation have been examined. Future work will involve using different metrics for comparison along with comparing different selections of text from the characters considered, along with the comparisons of translations of different authors by the same translator.

Notes

[1] See Farringdon (1996) for a detailed explanation of the origins of this technique, including detailed examples of the method's use in a legal setting.

[2] For the English versions of the plays, the print versions are collected in Ibsen, Archer, Aveling, Archer, and Archer (1890), Sharp's translations can be found in Sharp (1911), the collected works of Ibsen in German are to be found in Ibsen (1898) and the Norwegian collected works are found in Ibsen and Bull (1957)

[3] http://www.ibsen.net, last verified March 12, 2009, contains comprehensive information about Ibsen's life and work together with links to his plays in the original form and in translation.

[4] A Java based tool designed for this purpose, Lynch and Vogel (2007), describes the creation and benchmarking of this particular program.

[5] Hedda Gabler being the other one who springs to mind, further studies may incorporate a wider range of plays and characters.

[6] It is not fully clear from any forewords to the e-texts when exactly the translations themselves were first published, however it does state that the first performance in English was in 1890, using Archers translation, Sharp's translations were first published in 1911, according to http://www.leicestersecularsociety.org.uk/library_shelf.htm, last verified March 12, 2009

References

Burrows, J. (2002). The Englishing of Juvenal: Computational Stylistics and Translated Texts. *Style*, 36 (4), 677–699.

Farringdon, J. (1996). *Analysing for Authorship: A guide to the Cusum technique*. University of Wales Press.

Ibsen, H. (1898). *Henrik Ibsens sämtliche Werke in deutscher Sprache*. S. Fischer.

Ibsen, H., Archer, W., Aveling, E., Archer, F., & Archer,

C. (1890). *Ibsen's Prose Dramas*. W. Scott.

Ibsen, H. & Bull, F. (1957). *Samlede verker: hundreårsutgave*. Gyldendal.

Lynch, G. & Vogel, C. (2007). *Automatic Character Assignation. In Proceedings of AI-2007 Twenty-seventh SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 335–348.

Rybicki, J. (2006). Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations. *Literary and Linguistic Computing*, 21 (1), 91–103.

Sharp, R. (1911). *Henrik Ibsen, Ghosts and Two Other Plays*. J.M Dent.

Somers, H. & Tweedie, F. (2003). Authorship Attribution and Pastiche. *Computers and the Humanities*, 37 (4), 407–429.

Vogel, C. & Brisset, S. (2007). Hearing Voices in the Poetry of Brendan Kennelly. *Belgian Journal of English Language & Literature*, 1–16.

Vogel, C. & Lynch, G. (2008). Computational Stylometry: Who's in a Play?. In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*., pp. 189–194. Springer.

A Tool Suite for Automated TEI Encoding

Gerald C. Gannod
Miami University
gannodg@muohio.edu

Laura C. Mandell
Miami University
mandellc@muohio.edu

Holly L. Connor
Miami University
connorhl@muohio.edu

Abstract

The importance, benefits, and utility of encoding documents using markup guidelines such as TEI have long been recognized. However, the learning curve associated with using TEI has thus far inhibited widespread use. In this paper, we describe a tool suite that we have developed to facilitate adoption via the support for TEI in the ubiquitous Microsoft Word system. We discuss the motivation for such a tool suite and provide an overview of the primary capabilities.

Introduction

The Text Encoding Initiative (TEI) was established to provide a uniform set of guidelines for marking up a wide variety of text-based documents (Ide et al., 1995). The difficulties of learning TEI are formidable. The intended users of TEI are humanists, librarians and others who are interested in the long-term archival value of literary works. In contrast, the eXtensible Markup Language (XML), the framework upon which TEI is defined, was developed primarily by technologists as a way to encode data, documents, and other information into forms that are easily *readable by computers*. Clearly, the needs, backgrounds, and training of these two disparate communities do not directly coincide. Given the challenges inherent in learning XML and the encoding standards defined by the TEI, adoption of the TEI for encoding literature has been limited. Not only that, once project designers have learned it, they do not have the programming and/or scripting knowledge that they need in order to transform documents from TEI into the wide variety of possible uses for encoded documents, be they metadata files for export to libraries, repositories, and online search engines, or full-text documents in web pages and electronic books. This further reduces the current adoption of TEI while presenting an ever-ominous future where adoption of the standards will be made mandatory.

With these issues in mind, the goals of our research are two-fold. First, we are interested in developing a methodology for encoding literature that focuses on the end-user experience. By "enduser," we do not mean the users of archives nor readers of digitized texts. Rather, we mean the enduser of the tool suite, the faculty member using it to encode documents. Much of the encoding community (especially in regards to the TEI) has focused on the developer's experience in that the encoding of documents is biased towards convenience of writing and constructing post-processing tools that can parse and manipulate documents after they have been encoded. In contrast, we are interested in developing tools -in an environment and style familiar to the end-user audience: professors, researchers, and students in the Humanities. Second, we are interested in facilitating a wide breadth of end-user tasks. In addition to the encoding task, endusers are interested in transforming documents into many different forms, extracting meta-data, and forming queries to find interesting characteristics found in literature. Unfortunately, these tasks are not easy to learn and have to date required detailed programming knowledge.

In this paper, we describe tools that we are developing as add-ons to Microsoft Word to support the encoding task. As such, we realize many benefits. First, an overwhelming majority of users in the target community already use Word; as a result the learning curve associated with learning and using TEI may be reduced. Instead of having to learn a new program as well as TEI, the target user community will only have to learn new capabilities of Word. Second, the framework on which the tools are built is based on a platform (Microsoft Office) that is guaranteed to have some longevity. Thus end-users can be assured that the software will enjoy long-term support. Third, the programming model associated with Microsoft Word is as powerful as any other modern programming language. Thus software developed for the platform is only limited by what is possible in programming.

Background and Related Work

Encoding

SUNY University Press now requires authors not only to scan but also to run OCR on facsimile editions that SUNY has promised to publish. SUNY provides but one example of a general trend: publishers will be requiring authors to perform many of the operations that publishers were formerly responsible for, including perhaps XML encoding. Moreover, in one of the newer publishing models, the digital, print-on-demand model, Rice University Press will publish monographs submitted to NINES, the Networked Infrastructure for Nineteenth-Century Electronic Scholarship – a peer-reviewing organization for digital publications and an advocate for the use of the TEI guideline.

TEI has, since its inception in 1987, attempted to develop a set of elements (tags) adequate for describing every document that humanists, linguists, and librarians could imagine wishing to save for posterity. The first foray into trying to make TEI user-friendly involved developing TEI Lite and a system for teaching TEI. While TEI Lite provides a more limited tag set, one must still use an XML editor in order to use it – and, we would argue, it is not limited enough; it is still a set of elements designed for every imaginable kind of document.

Related Work

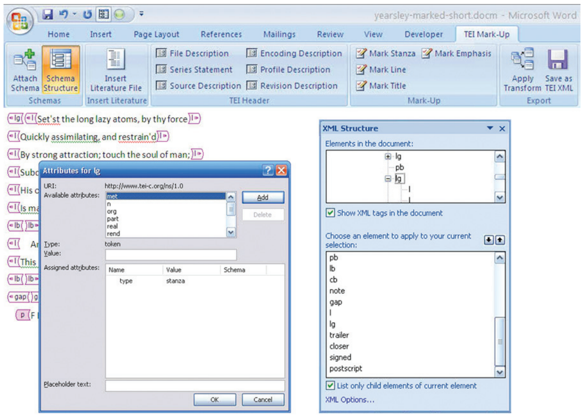
The TEI website maintains a list of many different tools that support the use of TEI markup (TEI Tools, 2008). In addition, there are numerous XML editors available on the market, the most popular being *XMetal, HTML Kit, Altova XMLSpy*, and *oXygen*. oXygen has been recommended by the TEI Consortium and is routinely used in TEI workshops. Recently, the oXygen company released another, easier version of its editing system called "oXygen Author" but still based on a direct XML manipulation model. While the tools are numerous, they are by-and-large focused on the editing of documents at the atomic level (e.g., tag-by-tag). The tools offer a lot of flexibility but require a great deal of knowledge about the encoding schemas. In addition, they do not preserve the original document formatting.

The Ajax XML Encoder (AXE) is a web-based tool that utilizes Ajax to support multi-user encoding of XML documents (Reside et al., 2008). The tool provides an interface that is accessible using a standard web browser to modify and manipulate XML documents. The tool also supports encoding of binary documents such as images and audio. As an alternative to the aforementioned XML editors, AXE facilitates *collaboration* between several users in an environment that is meant to be more accessible to common users.

Approach

Philosophy

Our tools have been built as an add-on for the 2007 edition of Microsoft Word which allows for XML editing and validation according to a W3 schema provided by the enduser. Thus, by giving users access to TEI tags in Word, we familiarize people with the notation and correct their uses via on-the-fly validation, thus teaching endusers about TEI while making those tags even more comfortable to use. By narrowing and correcting people's tag-choices and creating software designed to accomplish very specific general tasks, we give endusers a general understanding of TEI. We believe that they will be encouraged by this gentle learning curve to learn more about TEI. However, they can also simply pass their documents onto professional archivalists and TEI experts who can then perform deeper levels of coding and more advanced manipulations. In particular, we are seeking to achieve an 80/20 balance in the encoding of documents. That is, our goal is to support automatic encoding of 80% of a document while leaving the remaining (and more interesting) 20% of the encoding to an expert coder. The TEI Consortium has made the deliberate decision to ask scholars to learn a complex but rewarding coding system. Our tools provide the first step up into that system, and allow authors throughout the academy – not just experts in digital humanities – to contribute to developing the digital archive.



yearsley-marked-short.docm - Microsoft Word

Set'st the long lazy atoms, by thy force

Quickly assimilating, and restrain'd

By strong attraction; touch the soul of man;

Attributes for lg

URI: http://www.tei-c.org/ns/1.0

Type: token

Name	Value	Schema
type	stanza	

XML Structure

☑ Show XML tags in the document

☑ List only child elements of current element

Figure 1. Marked Up Document with Attribute Dialog and Structure View

Capabilities and Example

Our tool suite when viewed without visible tags looks like an ordinary Microsoft Word document. Figure 1 depicts part of a typical view that a user would encounter if using our system. The TEI Mark-Up tab reveals the supported tasks for encoding a poetry document. The user can choose to edit the document in a normal Word mode, or by selecting the "Schema Structure" button and checking the "Show XML tags in the document" checkbox, can reveal the TEI XML tags. The editor allows the user to browse the XML elements contained in the document and presents a list of a tags that can be inserted in the current context in a fashion similar to a view found in oXygen, as depicted on the right hand side of the figurThe real power of our approach is in the "Mark-Up" functions. For instance, in this version of the tool, if an entire block of text representing a stanza in a poem is highlighted, clicking on the "Mark Stanza" button will automatically tag the section with an enclosing "lg" tag, mark it with an attribute of "type = stanza", and encode every line in the stanza with either an "l" tag (in the case where there is text), or an "lb" tag (in the case of empty lines). By targeting these common "high payoff" encoding tasks, we are able to quickly encode a large majority of a document, thus freeing the encoder to focus attention on more interesting encoding tasks.

A number of functions are currently supported by our toolset including saving the marked up Word document into a raw TEI encoded XML file as well as the ability to apply the use of XSLT transformations on the encoded Word document. At the moment, we have add-ons that support mark-up of prose, poetry, and epistolary literary forms.

Our primary evaluation of this tool suite has come through applying the tool to real encoding tasks. We have used the technique to encode a large number of documents in a short period of time. Specifically, in a recent three-week period, we used the tool suite to encode an archive that contained hundreds of letters. By the time we demonstrate the tool at DH2009, we will have and so be able to present feedback from the students who used the tool as a way into understanding TEI markup.

Conclusions and Future Investigations

When compared with other XML editing tools, our Word add-ons offer similar capabilities; Word out of the box can be used as a fully functional XML editor, and our tool suite makes it into a fully functional TEI editor. While Word is of course proprietary, it is the word-processing ***program most often used by the target audience*** on both PCs and Macs. The tools will currently work only in the PC version of Word 2007. The supported programming model in Word (Visual Basic for Applications and C#.NET, etc.) provides a ***powerful suite of capabilities*** that enable construction of a wide variety of functions that support the encoding task. While Word does not support synchronous real-time collaboration on documents, it does support ***asynchronous collaboration*** through the "track changes" view. While programmers might want to use configuration management systems such as CVS and subversion, digital humanists who are comfortable with web forms and wiki or document management systems use tools such as Google docs. The endusers whom we target would find learning to edit in a wiki page an onerous distraction from the encoding task at hand. Furthermore, using a programmer-oriented configuration management system would pose enormous challenges beyond just learning to encode using TEI. Most of these users will already be familiar, however, with the reviewing and commenting tools in Word. As a result, our tool suite will allow multiple collaborators among traditional digital humanists to contribute to the encoding of a single document in a manner that allows them to track what changes were made and by whom.

Our future investigations include developing tools within Word that facilitate development of XSLT transformations as well as development of a general approach for automatically generating different encoding templates that support a wide variety of literary forms.

In tracking feedback from users, we will be interested in looking at whether and when they begin to consult the TEI P5 guidelines for more detailed tagging information, whether or when they switch from our Word Macros program to an XML editor such as oXygen, and whether or when they pass their documents onto experts for completion (that is, at what stage of coding). We already know that working directly with XML is the best way to learn it, but our product does not promise the shortest path. Rather, because the Word Macros making the learning curve for TEI most gentle at first, and steeper only later, if the enduser chooses to go on, we hope that the product gets people who need to encode documents and create digital archives the wherewithal to use TEI *at all*.

References

Ide, N. and Véronis, J. (1995). *Text Encoding Initiative: Background and Context*. Springer-Verlag.

Reside, D. and Lord, G. (2008). *AJAX XML Encoder*. On-line available at http://mith.umd.edu/mithresearch?id=19 (accessed November 8, 2008).

TEI Tools Page, (2008). TEI: Text Encoding Initiative. On-line available at http://www.tei-c.org/wiki/index.php/Category:Tools (accessed November 8, 2008).

Accessibility, Usability and the New Face of NINES

Dana Wheeles
University of Virginia
dw6h@cms.mail.virginia.edu

Laura Mandell
Miami University of Ohio
mandellc@muohio.edu

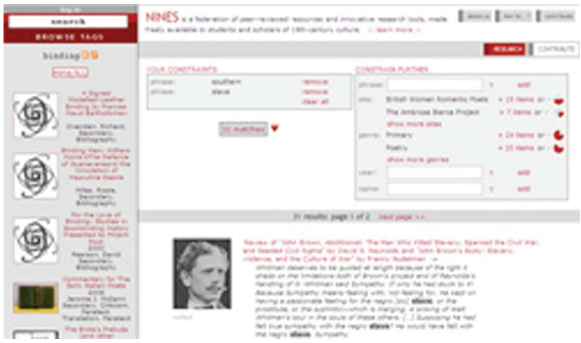
Nick Laiacona
Performant Software
nick@laiacona.com

NINES stands for a Networked Infrastructure for Nineteenth-century Electronic Scholarship, a scholarly organization in British and American nineteenth-century studies devoted to forging links between the material archive of the nineteenth century and the digital research environment of the twenty-first. In practice, these efforts have centered on building a portal that enables faceted searches of digital scholarly resources that we have peer-reviewed. This web portal also allows users to collect object and build exhibits online.

When NINES first launched its search interface in 2006, the institutional challenges of aggregating scholarly projects online were only the beginning. The NINES search and collect interface, powered by a **Coll**ecting and **Ex**hibiting tool called Collex, was a new thing, relatively speaking. Users were (and still are) very familiar with the idea of search engines: Yahoo and Google have long played a central role in any internet experience. But faceted browsers, or tools that allow the user to continually refine the categories of their search in order to pinpoint resources from large pools of data, were in their infancy. The most efficient of these were just beginning to be implemented in the commercial sector: Amazon.com is perhaps the best example of this. But the NINES team needed to adapt the power and flexibility of a faceted browser for scholarly research (rather than online shopping) and design it in a way that communicated the organization's reputability as a peer reviewer **and** a forward-looking software developer for humanities scholars. See the result below.

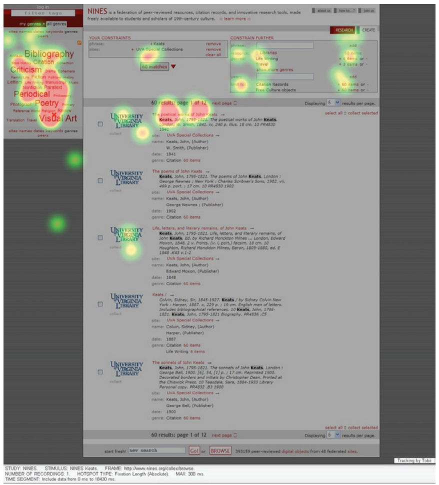
A left-hand sidebar doubled as a user tag cloud and a list of collected resources, while the central portion of the interface was dedicated to searching and browsing.

Searches in a faceted browser can become quite elaborate, so an area was needed to keep track of the constraints added by the user. The resources aggregated within NINES also needed ample screen real estate, as did the various genre designations of the objects for a different kind of browsing. Search results were listed below this frame, and were often "below the fold," requiring scrolling.



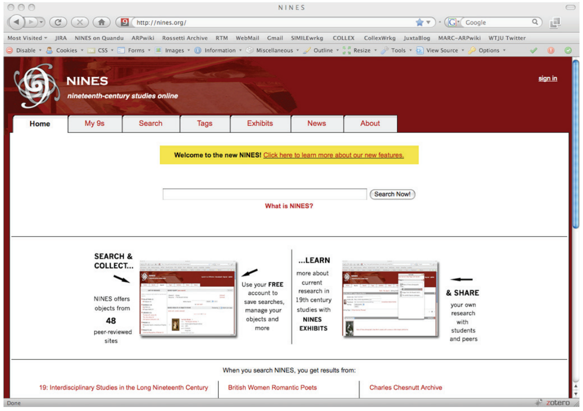
The NINES interface was a difficult thing to build, and, overall, it was a successful endeavor. Scholarly interest in NINES grew, and the number of digital objects leapt from roughly 70,000 to a whopping 300,000. Yet we found ourselves with very little feedback about the usability of this application. Were the targeted users (scholars and students of nineteenth-century studies) comfortable with NINES? Were they making use of the many features Collex provided? The development team decided to schedule a formal usability study to get answers.

As with any study of this kind, many of those answers were surprising. Five helpful participants in the NINES Summer workshop at Miami University demonstrated the strengths – and weaknesses – of the Collex interface.


Bibliography
Criticism
Periodical
Poetry
Visual Art
UNIVERSITY of VIRGINIA LIBRARY
60 results: page 1 of 12 next page
STUDY: NINES STIMULUS: NINES Keats FRAME: http://www.nines.org/collex/browse
NUMBER OF RECORDINGS: 1 HOTSPOT TYPE: Fixation Length (Absolute) MAX: 300 ms
TIME SEGMENT: Include data from 0 ms to 10430 ms
Tracking by Tobii

As can be seen in this screenshot demonstrating the eye movement during one of the usability sessions, the user's focus is scattered at any given moment. The tag cloud at the left was an attractive feature, but its presence alongside search results put it in direct competition with them. Users frequently jumped back and forth from the tag cloud to the summary of the search results and the constraints. But once they found objects, users did not think of creating an account to collect those objects until they were prompted to do so by the test, and even then, creating a free account gave them pause. Although each person commented favorably on the variety and amount of information integrated into NINES, every single person tested said that the interface was a mystery – this after having spent a full 20 minutes using it.

After some consideration of the results, we decided that the fundamental structure of Collex was sound: its features were compelling and the searching capabilities attractive. In a way, Collex was *too* powerful, to the extent that it was intimidating. Everything about our interface, from the Home page to the Collex engine, had been conceived in scholarly and theoretical terms, and demanded a similar dedication from the user. And while NINES has always sought to appeal to the growing community dedicated to the mission of excellence in digital scholarship, we never intended to limit our users exclusively to its membership. We wanted Collex to be a friendly environment: a place where both scholars and their students go to conduct research in the nineteenth-century studies online.



Welcome to the new NINES! Click here to learn more about our new features.

SEARCH & COLLECT... NINES offers objects from 48 peer-reviewed sites

Use your FREE account to save searches, manage your objects and more

...LEARN more about current research in 19th century studies with NINES EXHIBITS

& SHARE your own research with students and peers

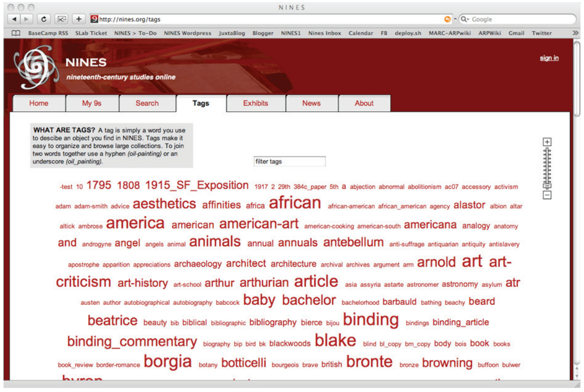
When you search NINES, you get results from:

19: Interdisciplinary Studies in the Long Nineteenth Century | British Women Romantic Poets | Charles Chesnutt Archive

As we began to sketch out the Collex redesign, our primary goal was to break up the Collex interface into its component features, giving each one its own important area of the site. In many ways, our task was much simpler than the one that faced the original designers of Collex – in the intervening years, the number of faceted browsers had multiplied considerably leading to the emergence of web conventions for their styling and organization. Our 'market research' was directed to what worked and what didn't among the giants of internet commerce: Amazon's browser was the most intuitive, but Ebay's method of displaying search constraints was more informative. And, it must be remembered, a scholar is not always looking for that one perfect object, as is the online shopper, but rather a collection of objects that fit ever more specific requirements. Collex needed to allow users to tunnel into our data quickly and efficiently, all the while reminding them of the numerous other materials available.

The new NINES home page was designed to communicate two things: first, that this is a scholarly organization, and second, that searching is an integral part of the site. Making use of the familiar Google search blank, we hoped to eliminate our site's high bounce rate (the number of visitors who leave almost immediately after arriving) and invite those with merely casual interest to stick around and explore.

The tag cloud – one of the most forward-looking and dynamic parts of the NINES interface – was given its own page, with plenty of room for browsing the interests of NINES users.


NINES
http://nines.org/tags
Google
BaseCamp RSS SLab Ticket NINES > To-Do NINES Wordpress JuxtaBlog Blogger NINES1 Nines Inbox Calendar FB deploy.sh MARC-ARPwiki ARPWiki Gmail Twitter
NINES
nineteenth-century studies online
sign in
Home
My 9s
Search
Tags
Exhibits
News
About
filter tags
-test 10 1795 1808 1915_SF_Exposition 1917 2 29th 384c_paper 5th a abjection abnormal abolitionism ac07 accessory activism
adam adam-smith advice aesthetics affinities africa african african-american african_american agency alastor albion altar
altick ambrose america american american-art american-cooking american-south americana analogy anatomy
and androgyne angel angels animal animals annual annuals antebellum anti-suffrage antiquarian antiquity antislavery
apostrophe apparition appreciations archaeology architect architecture archival archives argument arm arnold art art-
criticism art-history art-school arthur arthurian article asia assyria astarte astronomer astronomy asylum atr
austen author autobiographical autobiography babcock baby bachelor bachelorhood barbauld bathing beachy beard
beatrice beauty bib biblical bibliographic bibliography bierce bijou binding bindings binding_article
binding_commentary biography bip bird bk blackwoods blake blind bl_copy bm_copy body bois book books
book_review border-romance borgia botany botticelli bourgeois brave british bronte bronze browning buffoon bulwer

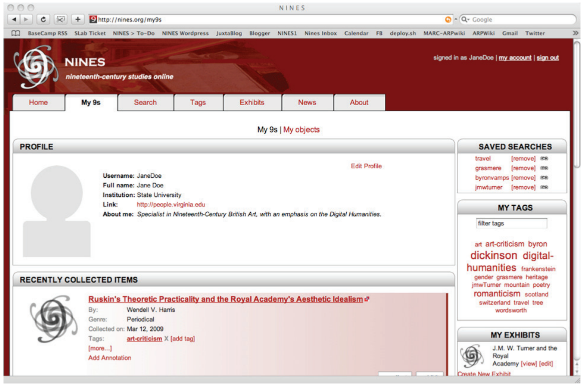
WHAT ARE TAGS? A tag is simply a word you use to descibe an object you find in NINES. Tags make it easy to organize and browse large collections. To join two words together use a hyphen *(oil-painting)* or an underscore *(oil_painting)*.

Tool tips and help text were added ("What are tags?") whenever the use of features was subtle or required specialized knowledge of social media software. And finally, a space was reserved for a blog, so that users could appreciate NINES as an active and ever-growing institution, rather than a static website or finding aid.

But as these refinements were implemented, we kept running into a logical disconnect between the kinds of operations one could do in NINES. Searching and browsing are available to any and everyone who visits the site, and are easily found in the main navigation tabs. However, what makes NINES unique is that is also a scholarly workspace: after setting up an account, you are able to collect objects, tag them, and re-mix them into a essay or "Exhibit" of your own. NINES is first and foremost conceived as a community for scholars, a place where one can simultaneously contribute to and benefit from an ongoing discourse on nineteenth-century studies. Unfortunately, even with the new design, this aspect of our mission was still not clear.

Taking a cue from Facebook, Flickr and other social media sites, the NINES team conceived the "My 9s" page, each user's private homepage within NINES.

On this page, all your authoring, editing and collecting efforts are centralized: as a user you have control over how much of your personal information is shared with other users while your tags, saved searches and recently collected items are easily accessed. Exhibits, which had previously been a mysterious and intimidating option are now offered as a logical outcome of all your work within NINES – why not share some of that hard work with your peers? By enlarging the (pre-existing) social networking attributes of NINES, we were able to fully demonstrate the power and utility of Collex, all the while making the many operations possible within it appear more manageable and enticing.


NINES
http://nines.org/my9s
Google
BaseCamp RSS SLab Ticket NINES > To-Do NINES Wordpress JuxtaBlog Blogger NINES1 Nines Inbox Calendar FB deploy.sh MARC-ARPwiki ARPWiki Gmail Twitter
NINES
nineteenth-century studies online
signed in as JaneDoe | my account | sign out
Home
My 9s
Search
Tags
Exhibits
News
About
My 9s | My objects
PROFILE
Edit Profile
RECENTLY COLLECTED ITEMS
SAVED SEARCHES
travel [remove]
grasmere [remove]
byronvamps [remove]
jmwturner [remove]
MY TAGS
filter tags
art art-criticism byron dickinson digital-humanities frankenstein gender grasmere heritage jmwTurner mountain poetry romanticism scotland switzerland travel tree wordsworth
MY EXHIBITS
J.M. W. Turner and the Royal Academy [view] [edit]
Create New Exhibit

Username: JaneDoe
Full name: Jane Doe
Institution: State University
Link: http://people.virginia.edu
About me: *Specialist in Nineteenth-Century British Art, with an emphasis on the Digital Humanities.*

Ruskin's Theoretic Practicality and the Royal Academy's Aesthetic Idealism
By: Wendell V. Harris
Genre: Periodical
Collected on: Mar 12, 2009
Tags: art-criticism X [add tag]
[more...]
Add Annotation

Going forward, we hope to incorporate other social networking protocols to encourage the growth of a community of NINES scholars. The "My 9s" page is an initial step in this direction, and we are also planning a NINES Facebook application, which will allow us to leverage an existing social network to form our own. Because of the way NINES is structured, we are only as strong as the community that we are able to build. Part of this process involves peer-review and aggregation via RDF metadata; our summer workshops for new projects are another important activity in this regard. But our web interface and software tools are central to the ways that NINES will gather scholars together. Through a combination of usability studies, development meetings, discussions with scholars in the field, and pure trial-and-error, we have come to recognize the importance of re-tooling, an ongoing process that is constantly engaged with the social and technical ecology of the web.

Visualization and Landscape in the Digital Humanities

John Melson
Brown University
John_Melson@brown.edu

What does "landscape" mean for digital humanities? In the fields of traditional literary and cultural studies, representations of real and imagined landscapes have routinely been scrutinized as windows into particular cultural moments, spaces in which complex relationships between humans in their natural and constructed environments are worked out and assigned social, political, and cultural meaning. Digital humanities have developed a rather different sense of the term, however. Maps of the publication and circulation of printed texts, spatial visualizations of historical urban plans, three-dimensional reconstructions of archeological sites: all these and more have now attained the status of "landscapes" in the digital humanities, for they situate texts and cultural artifacts in space and time. Yet even as such technology-driven approaches continually reshape the "humanities research landscape" (to use the term in a different sense) they also raise new questions about the kinds of intellectual activities they promote, testing what Martyn Jessop has recently called the "dynamic process" of creating knowledge (Edmond 2005; Jessop 2008).

This paper explores the changing significance of "landscape" as a keyword in the humanities, one whose relevance is amplified dramatically by the turn toward technologies of visualization and the development of theoretical and methodological frameworks for integrating such technologies into more conventional forms of textual scholarship. Using as a test case a small set of eighteenth-century Anglophone texts that theorize landscape aesthetics and deploy tropes of real and imagined landscapes to reflect on contemporary cultural relations across the Atlantic, I argue that methods for representing some of their salient features in visual rather than textual forms suggest new ways of understanding cultural exchange in a historical moment characterized by intense anxieties about the temporal and spatial distances separating England from its North American colonies during the "British diaspora" (Tennenhouse 2007). This claim, and the specific visual representations that support it, forms the basis for considering how models of textually-mediated cultural authority were themselves constructed at the nexus of space and time represented by landscape aesthetics. This latter claim extends what might initially seem to be a straightforward literary and historical argument to the present moment in digital humanities. Examining these cultural constructions through the lens of digital visualizations affords insight into the present-day relationship between scholars' interpretive acts and the cultural weight we grant such activities when they happen in and through digital media.

In 1690, a short tract titled *The Geometry of Landskips and Paintings Made Familiar and Easie*, published in London, noted the artifice involved in landscape representation: "The Geometry of Painting is rather Optick or Perspective, than real…A Landskip is therefore a rather neat contraction or Epitome of things visible than a real view of them." Though somewhat cryptic, the anonymous pamphleteer's opening remarks expose the carefully crafted fiction that landscape paintings represent real places as they actually are; to the contrary, he insists, landscapes are deliberately compressed expressions of an idealized, imaginary world—the world as we would like to see it. In this regard, the pamphlet's frank admission of the ideal landscape's constructedness confirms the common view among literary scholars and cultural historians that textual and visual landscapes have historically naturalized ideologies of power and authority within the contours of real and imagined geographic space (Barrell 1972, 1980; Bermingham 1986; Williams 1973).

Yet the pamphlet's titular promise to make landscape "Familiar and Easie" resonates also with this paper's examination of the present-day digital humanities "landscape," and particularly the role of visualization in advancing both close and "distant" reading practices. With calls in some quarters for "a new object of study…in which the reality of the text undergoes a process of deliberate reduction and abstraction," and with the increasing conviction that text visualizations furnish that object, the question of graphical and visual literacy has become all the more pressing (Moretti 2000). As Jessop notes, "Humanists… have little in the way of visual literacy," an observation which leads him to conclude that collaboration between digital humanities projects and artists is necessary in recognition of the fact that "aesthetics are deeply embedded in the effective use of the medium of digital visualization" (Jessop 2008). That visual literacy is no more "familiar" or "easy" for many now than it was at the end of the seventeenth century suggests that the historical evolution and function of landscape has much to teach us about the current promises and pitfalls of humanities visualization.

It is on this point that my paper's examination of visualizing patterns of reference in a small set of eighteenth-century Anglo-American texts advocates the usefulness of thinking in broader terms about “the digital” as a landscape in its own right—a landscape that exposes rather than conceals its own constructedness and, in so doing, reveals new modes of legitimating and authorizing the scholarly activities to which it belongs. In other words, by regarding visualizations as composing another kind of landscape it becomes possible to ask how digital humanities, in the process of collecting, shaping, and presenting textual data, themselves participate in what amounts to a regime of aesthetic judgment. Like the landscape aesthetics of the eighteenth century, the interpretive lexicon authorized by digital visualization raises new questions about scholarly authority, models of judgment, standards of evidence—indeed, questions not only of how we read and interpret texts but also who is authorized to read them in particular ways. The idea of landscape has historically been deeply involved in negotiating the transmission of cultural authority. Its renewed formal significance for scholars grappling with the relationship between quantitative textual or spatial data and the “fuzziness” of highly subjective interpretive practices (to mention but one scenario) in which, once again, authority is stake, is therefore highly germane to the methodologies that inform scholarly practice at the intersection of digital and non-digital humanistic inquiry.

References

Anonymous. (1690). *The Geometry of Landskips and Paintings Made Familiar and Easie*. London: Richard Baldwin.

Barrell, J. (1980). *The Dark Side of the Landscape: The Rural Poor in English Painting, 1730-1840*. Cambridge: Cambridge University Press.

Barrell, J. (1972). *The Idea of Landscape and the Sense of Place, 1730-1840: An Approach to the Poetry of John-Clare*. Cambridge: Cambridge University Press.

Bermingham, A. (1986). *Landscape and Ideology: The English Rustic Tradition, 1740-1860*. Berkeley: University of California Press.

Edmond, J. (2005). The Role of the Professional Intermediary in Expanding the Humanities Computing Base. *Literary and Linguistic Computing*, 20: 367-380.

Jessop, M. (2008). Digital Visualization as a Scholarly Activity. *Literary and Linguistic Computing*, 23: 281-293.

Moretti, F. (2000). *Graphs, Maps, Trees: Abstract Models for a Literary Theory*. New York: Verso.

Tennenhouse, L. (2007). *The Importance of Feeling English: American Literature and the British Diaspora, 1750-1850*. Princeton: Princeton University Press.

Williams, R. (1973). *The Country and the City*. New York: Oxford University Press.

Modernist Magazines Project

Frederico Meschini
De Montfort University
fmeschini@dmu.ac.uk

The critic Michael Levenson warned that "A coarsely understood modernism is at once an historical scandal and a contemporary disability". The Modernist Magazine Project aims to refine and enhance the record through the production of a scholarly resource and comprehensive critical and cultural history of modernist magazines in the period 1880-1945. So-called 'little magazines' were small, independent publishing ventures committed to new and experimental work. Literally hundreds of such magazines flourished in this period, providing an indispensable forum for modernist innovation and debate. They helped sustain small artistic communities, strengthened the resolve of small iconoclastic groups, keen to change the world, and gave many major modernists their first opportunities in print. Many of these magazines existed only for a few issues and then collapsed; but almost all of them contained work of outstanding originality and future significance.

The project aims to document and analyse the role of both fugitive and more established magazines and to consider their contribution to the construction of modernism in Britain, Europe and North America. It will result in a 3 volume Critical and Cultural History of Modernist Magazines, an Anthology and an online resource, comprising an index of magazines, bibliographical and biographical data, selected contents and web links.

What this paper wants to focus on is the design and actual implementation of the online resource together with all the critical and technical thinking involved in defining every step of workflow.

The first part involved thinking about the proper formats and standards which would best suit the project needs and at the same time would be able to be conformant with the current best practices and the widely advocated interoperability and preservation issues. Therefore a natural choice has been the adoption of Library of Congress standards related to the implementation of both general descriptive metadata and specific solutions for the domain of digital libraries. In particular the standards involved are MODS[1], METS[2], MADS[3], having also a look at PREMIS[4].

All these standards at the same time provide for and prefer an XML serialization, which, given the almost ubiquitous support of this syntax by the current web frameworks, allows for many different possibilities in choosing the underlying technology. The final choice consisted in the eXist XML Native Database[5]. Being under open-source development for several years, this database has been constantly growing in features, performance and user base, becoming not only a mere data storage layer, but, thanks to the integration with Apache Cocoon[6] which allowed for the possibility of creating processing pipelines composed by XQuery and XSLT, an actual and complete web framework.

A particular focus will be dedicated to the strategies implemented for converting from non structured data to MODS metadata, so to facilitate the work of the developer, since the data collecting has been made by an expert in modernism, but with few skills in metadata management. A key point has been the preparation of Excel forms, which allowed for the gathering of semi-structured data and the subsequent development of a conversion program based on Apache POI[7] and JDOM[8] API for the direct creation of MODS compliant metadata. In particular the difference between the tabular structure of Excel and the nested tree structure of XML required the implementation of some particular algorithmic strategies.

What will be discussed next is the use of MADS for recording the different name versions and pseudonyms, the creation of a MODS extension, compatible with METS, for recording the information about the digital images of the magazines pages, and the subsequent display of these images using the YUI framework[9].

The last part of the presentation will be about a proof-of-concept integration between the Modernist Magazines Project and the thematic-related Modernist Journal Project, based on the use Web 2.0 techniques, in particular Ajax and REST Web Services[10].

Notes

[1] *Metadata Object Description Standard* <http://www.loc.gov/standards/mods/>.

[2] Metadata Encoding & Transmission Standard <http://www.loc.gov/standards/mets/>.

[3] *Metadata Authority Description Schema* <http://www.loc.gov/standards/mads/>.

[4] *Preservation Metadata Maintenance Activity* <http://www.loc.gov/standards/premis/>.

[5] *eXist* <http://exist-db.org/>.

[6] Cocoon <http://cocoon.apache.org/>.

[7] <http://poi.apache.org/>.

[8] *<www.jdom.org/>.*

[9] *The Yahoo! User Interface Library <http://developer.yahoo.com/yui/>.*

[10] Costello, Roger L. Building Web Services the REST Way. <http://www.xfront.com/REST-Web-Services.html>

Should an electronic edition walk, swim or shake its tail feathers?

Frederico Meschini
De Montfort University
fmeschini@dmu.ac.uk

The scholarly activity of creating a critical edition of a literary work is an extremely complex process, composed of many steps, each involving different features and therefore requiring different skills. For instance Wilhelm Ott described eight different steps, each one corresponding to a particular software module in the TuSTEP system, starting with the collection of witnesses and ending with the edition publication, passing through intermediate phases such as collation, constitution of copy-text, compilation of apparatuses and indexes creation (Ott, 1992) .

But it would be very naïve to suppose that the current paradigmatic and epistemological shift from print to electronic medium in producing scholarly editions would be without consequences for textual editing as recently pointed out by Peter Shillingsburg (Shillingsburg, 2006). Whereas the printed publication is the last and final step in Ott's workflow, in the digital world this same phase is no longer a dead end, allowing for further, and potentially endless processing.

The actual crux is the analysis, definition and understanding of these consequences, which, to be as possible comprehensive and effective as possible should be carried out both on an intellectual and a practical level. The more tangible effects are focused at the two extremes of the scholarly process, which from being the most mechanical steps now acquire a new dimension and importance: on the one hand the transcription of the textual artefacts, which at present is a structural and semantic encoding, and, on the other, the dissemination modalities, which address how the edition is initially assembled and subsequently published. While much theoretical discussion has focused on text encoding, the same is not true, with some notable exceptions, for the other extreme: the actual creation of the edition.

An implicit complexity seems to be inseparable from any thinking or talking about electronic scholarly editions. It's not by chance therefore that Susan Hockey in *Electronic Texts in the Humanities* wrote that "Much confusion seems to surround the topic of electronic editions" (Hockey, 2000). This is due to the fact that the digital edition is a dynamic and mutable object in itself: its nature as an electronic text which, like spoken language, allows it to be processed in a reflexive way, and therefore being augmented with new features and uses. But there is more than this. The extreme complexity of critical digital editions is generated by an aspect almost always neglected but at the same time quite fundamental: the multidimensionality of this strange animal, commonly known as 'electronic edition'. Both Jerome McGann (McGann, 2004) and Claus Huitfeldt (Huitfeldt, 1994) underline how the object 'text' has more than one dimension: why can the same principle not be applied to the environment which contains, preserves and allows the interactions with this object?

A first method of describing this complexity can be found in 'A framework for information system architecture' by J. A. Zachman (Zachman, 1987) where the author theorizes a possible formal structure for investigating the architecture of information systems. This structure is an interpretational matrix where the different perspectives, tied to the different user roles of the system, are joined by the different descriptions of the information system, each one referring to a particular model: functional, informational, technological, etc. The primacy conclusion is that an information system is represented by a whole set of architectural representations, each one with a particular nature, and communication between these different levels is a key issue. Scholarly editions can be considered as a specialized subset of information systems, if not very refined 'knowledge systems' at all, the main difference being that they respond to specific needs of humanities research and not to generic business needs. Therefore Zachman's proposal can also be applied to good effects in this particular field.

A further step in trying to formalize digital editions is by means of specialized frameworks expressly created for the digital library world. Even though electronic editions and digital libraries are two different paradigms, like men in Shakespeare's *Tempest* they share the stuff that 'dreams are made on'. The two frameworks are the 5S (Gonçalves et al., 2004) and DELOS (Candela et al., 2007). The 5S model is based on set theory and linear algebra, and using five primitive concepts (Stream, Structures, Scenarios, Spaces and Societies) it builds upon these a series of definition, thus being able to define what a digital library is without ambiguity. On the other hand the DELOS approach shares many similarities with the CIDOC-CRM ontology (Crofts et al., 2007), being based on an object-relationship model. With the help of these two models, I will analyze the differences between digital libraries and electronic editions.

Starting from an empirical analysis of concrete cases, some basic principles will be presented using a polarity approach. An actual limitation in an edition will be used as the starting point to develop an opposite principle which will be used to overcome it. These principles are the following: incompatibility vs. semantic umbrella/glue; sonic screwdriver vs. lego-block; blob vs. crystal snow; incompleteness vs. extensibility.

Finally the relationship between the latest innovations in electronic (web) publishing and scholarly editions will be examined. Using the "swimming" metaphor, Charles Michael Sperberg-McQueen pointed out the main difference between printed and electronic editions: while the former are embedded with some kind of implicit knowledge, the latter can contain the same knowledge expressed in formal languages, being moreover endowed with some active features (Sperberg-McQueen, 2002). This same difference between facts on one side and features on the other can be mapped to the current paradigms of Semantic Web and Web 2.0. An electronic edition is therefore an ideal place of interaction between these two different aspects of the WorldWideWeb, which are usually considered in opposition to each other.

References

Candela, L., D. Castelli, N. Ferro, Y. Ioannidis, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, M. Dobreva, V. Katifori, and H. Schuldt (2007). *The DELOS Digital Library Reference Model. Foundations for Digital Libraries Version 0.98*. DELOS Network of Excellence on Digital Libraries.

Crofts, N., M. Doerr, T. Gill, S. Stead, M. Stiff (eds) (2007) *Definition of the CIDOC Conceptual Reference Model: Version 4.2.2*. ICOM - International Council of Musesum.

Gonçalves, M. A., Fox, E. A., Watson, L. T., & Kipp, N. A. (2004). Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries, *ACM Transactions on Information Systems (TOIS)*, 22 (2): 270-312.

Hockey S. (2000). *Electronic Texts in the Humanities: Principles and Practice*. Oxford: Oxford University Press.

Huitfeldt C. (1994). Multi-dimensional texts in a one-dimensional medium, *Computers and the Humanities*, Vol. 28, No. 4-5: 235-241

McGann J. (2004). Marking Texts of Many Dimensions. In S. Schreibman, R. Siemens, J. Unsworth (eds.), *A*

Companion to Digital Humanities, Oxford: Blackwell, pp. 198-217

Shillingsburg, P. (2006) *From Gutenberg to Google: Electronic Representations of Literary Texts*, Cambridge University Press

Ott, W. (1992). Computers and Textual Editing, In Butler C. S. (ed.), *Computers and Written Texts*. Oxford: Blackwell, 1992. 205-226.

Sperberg-McQueen, C. M. (2002). How to Teach Your Edition How to Swim <http://www.w3.org/People/cmsmcq/2002/cep97/swimming.xml>

Zachman, J. A (1987). A framework for information systems architecture, *IBM Systems Journal*, 26 (3): 276-292.

Literature, "The Literary," and the Dataworld

Stuart Moulthrop
University of Baltimore
smoulthrop@ubalt.edu

This paper addresses a recurrent (perhaps constitutive) concern in electronic literature: as technical practices and affordances change, how should we distinguish between what is and is not *literature?* Liberal-minded critics like N. Katherine Hayles and Alan Liu have proposed a category called "the literary," a kind of cosmic cloud orbiting the stellar core of traditional literature. Some demur from this view, defending formal distinctions between literary and non-literary production (John Cayley), consonant with a neo-formalist movement in art and design (Ellen Lupton).

I trace this tension to a central dynamic of interactive media: the shift from CONTENT to DATA, which arises from radical change in the technical substrate of writing, where computation replaces inscription. Our technologies of writing and related sign-exchange no longer operate by containment in static, material structures, but now flow as impulses through processors and networks. The consequences of this shift are far from trivial or incidental. Containment yields to circulation, publishing and broadcasting to Web services. Where CONTENT was, we now find DATA, meaning *that which is given*, always in multiple dimensions:

- given IN to a system of encoding;
- given OUT through a system of communication;
- given UP to processing and transformation;
- given BACK as riff, recursion, and feedback.

Until fairly recently, the full impact of this change has been elusive, even in purportedly forward-looking circles. Generally speaking, electronic literature has engaged my sense of *data* in only limited ways. The early phase of electronic literature largely featured translation of lyric or narrative content-objects into data -- as in all my work to date. True, constructs like interactive fictions or procedural poems convert words into digital information, and evoke a text according to logical processes; but with the crucial exception of MUDs and MOOs, first-generation works generally do not involve multiple participants, or allow the core code to change with use.

Considering that constantly connected, high-speed digital networks have only been widely available for about a decade, these limits are not surprising.

With the arrival of such affordances around the turn of the century, digital culture has crossed a watershed. However much bemoaned by scholars, Wikipedia and its cousins have vastly increased the amount of not-entirely-erroneous information ready to hand. In the creative sphere, electronic writers and net artists have begun to exploit the content-data shift in more socially complex ways. Work by Jim Carpenter, Noah Wardrip-Fruin, Mark Marino, Jason Nelson, Kate Armstrong, and others has opened electronic poetry and prose to RSS feeds, Googlemaps, Twitter, FaceBook, and a range of other social information tools. These experiments complicate demarcations between writing and performance. Notably, they also generate the anxieties about literature and "the literary" that are the main subject of this paper.

As a long-time member of electronic literature's Golden Age team, my sympathies are divided. I like the new voices and believe deeply in the direction of their art, but at the same time I respect those who are more skeptically inclined. Like Cayley, I have my own investments in stable and salient expression, and would probably place certain limits on the social life of information. Like Lupton, I believe the explosion of creative energy enabled by emerging technologies should be shaped by sound understanding of formal principles.

So, as the sons of middle children will, I call for accommodation. However much we deal in data, the idea of content seems ineradicable. So in the rapidly evolving space of electronic media, literature and "the literary" must find ways to get along. While the sentiment alone is not worth much, I also offer some practical suggestions:

1. Anyone concerned about electronic literature (or indeed, digital culture more generally) must embrace *green technology* -- taking this adjective not in its popular, environmentalist sense, but in its other register of novelty or neoteny. Obedient to the Law of Mo(o)re, we should never expect our tools, or our scope of work, to attain the long-term stability of print culture, at least not in present generations. We paint in the rain.

2. Accordingly, we need to shape an aesthetic that embraces both that which is remembered and resonant, and that which is evanescent or transient. As Thomas Pynchon's heretical ex-Puritan ancestor argued, salvation is impossible unless some are passed over. Some child is always left behind. From which it follows that oblivion or preterition is an inverse grace. The stars cannot shine without those clouds of dust.

3. We need better ways to transfer understanding and technique: occasions for radicals of the dataworld to earn recognition from those of us still consecrated to content. Likewise, we could use forums in certain old hands interested in crossing the watershed might learn new approaches from those who have come after and gone ahead.

The Electronic Literature Organization and its international peers have all taken important steps toward these goals, and I will end by looking at a few of these developments.

Cosine Distance Nearest-Neighbor Classification for Authorship Attribution

John Noecker Jr.
Duquesne University
jnoecker@gmail.com

Patrick Juola
Duquesne University
juola@mathcs.duq.edu

There is a significant overhead cost when using techniques like support vector machines on large data sets. It would certainly be convenient to be able to use a less involved technique, but this often comes with a cost of less desirable performance. This results in a trade between time and memory restraints and prediction accuracy. It seems that a sophisticated technique like support vector machines must surely outperform a simple nearest neighbor classification. Fortunately, some recent results suggest that in fact a simple nearest neighbor classification using the normalized dot product (the so-called 'cosine distance') as a 'distance' performs comparably to radial basis function support vector machines for the task of authorship attribution. In some cases, this cosine distance classification actually outperforms SVMs.

Whether or not a space is linearly separable can have important consequences on performing classification within that space. An n-dimensional space containing two classes of points is said to be linearly separable if there exists an n-1 dimensional hyper-plane which separates the classes. A linearly separable space has the advantage that simpler classification will work for classifying points within that space. A simple distance metric may be sufficient to distinguish between two classes in a linearly separable space, while a more complex method like support vector machines or a neural network is necessary to capture nonlinear class boundaries. The primary advantage of linear separability is that is allows us to develop classification algorithms that are less computationally intensive. That is, instead of taking hours or even days to model the class boundaries, we can use simple algorithms which will achieve comparable results in only a few minutes. Linear classifiers tend to scale considerably better than their more complex counterparts. This is especially important when working with very large corpora, where training a support vector machine could take several days, while evaluating the cosine distance between documents in the corpus may take less than an hour.

We intend to present recent results in the field of authorship attribution suggesting that the normalized dot product nearest neighbor classification method is is comparable to radial basis function support vector machine classification methods. For this experiment, we made use of the Java Graphical Authorship Attribution Program (JGAAP-*www.jgaap.com*), a freely available Java program for performing authorship attribution created by Patrick Juola of Duquesne University. This modular program breaks the task of authorship attribution into three subtasks, described as 'Canonicization', 'Event Generation' and 'Statistical Analysis'. During the Canonicization step, documents are standardized and various preprocessing steps can occur. For this experiment, we have used a variety of combinations of three preprocessing steps: 'Strip Punctuation', 'Unify Case' and 'Normalize Whitespace'. Although the choice of canonicizers had some effect on the overall performance of the statistical analysis methods, the choice of preprocessors did not significantly affect the results. For the feature sets, we used characters, character bigrams, word lengths, word bigrams and words. We performed the experiments both with the full feature sets and with only the 50 most common features. For the statistical methods, as previously discussed, we used both radial-basis function SVMs and the normalized dot product scoring for nearest neighbor classification. We made use of the libSVM package for our SVMs.

In order to test this experiment on real world data, we have used the Ad-hoc Authorship Attribution Competition (AAAC) corpus. The AAAC was an experiment in authorship attribution held as part of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities. The AAAC corpus provides texts from a wide variety of different genres, languages and document lengths, assuring that the results would be useful for a wide variety of applications. The AAAC corpus consists of 98 unknown documents, distributed across 13 different problems (labeled A-M). An analysis method's AAAC score is calculated as the sum of the percent accuracy for each problem. Hence, a AAAC score of 1300% represents 100% accuracy on all problems. This score was designed to weight both small problems (those with only one or two unknown documents) and large problems equally. Because this score is not always sufficiently descriptive on its own, we have also included an overall accuracy rate in our experiment. That is, we calculate both the AAAC scoring and the total percentage of unknown documents which were assigned the correct authorship labels. These two scores provide a fair assessment of how the technique performed both on a per-problem and per-document basis.

In those cases where all events were included, the cosine distance classification actually outperformed radial basis function SVMs, while performing only slight worse on the most common event sets. This leads us to the conclusion that although much of the information necessary for authorship attribution is contained within the 50 most common events, it is the less common events that actually result in a sort of empirically linearly separably clustering of the data points. That is, when presented only with the 50 most common events as a feature space, we require a more complex classifier to model the class boundaries between different authors. However, as we increase the number of dimensions of this space by adding the less-frequently used features, it becomes possible to model these boundaries with only simple linear classifiers. Hence there is some important information contained even within the rarely occurring events in the feature set.

The finding that the normalized dot product performs comparatively to support vector machines is important because it is very simple and computationally tractable even in high-dimensional spaces. It is much quicker to calculate the dot product between two vectors than it is to train a support vector machine. Using a normalized dot product scoring nearest neighbor algorithm will allow very large data sets to be processed much more quickly and does not seem to cause a significant loss of accuracy. Hence, we propose that for some applications, the normalized dot product scoring may be an acceptable substitute for using a support vector machine.

Mashing Texts: Supporting collections level text analysis

Piotr Organisciak
University of Alberta
organisc@ualberta.ca

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

Stan Ruecker
University of Alberta
sruecker@ualberta.ca

Susan Brown
University of Guelph
sbrown@uoguelph.ca

Kamal Ranaweera
University of Alberta

Introduction

At the 2005 Summit on Digital Tools in the Humanities the need for tools for the Exploration of Resources [1] was identified as one of four opportunities for humanities computing tools. As a critical mass of evidence useful to humanities research becomes available on the web, researchers need tools for gathering the resources they need to ask questions, assembling and editing the evidence into study collections, and then analyzing those collections. This paper will discuss the Mashing Texts project [2] that has followed a persona usability design process to develop stories and a prototype for how a collections analysis tool might work to support humanities research. In particular, we will:

1. Demonstrate the JiTR (Just-in-Time Research) prototype and how it can be used to assemble a collection, edit items, and analyze them

2. Discuss the persona design process and the stories we generated to describe potential use of such a tool. In particular we will outline two different communities of users that ran through the design process.

3. Discuss the technical and service implications of such a tool which has to integrate with digital collections, digital archives, and text analysis systems.

We believe the humanities computing community can make the case for such an environment and that we can outline models for integration so that such a tool can interoperate with the content archives and text analysis tools available.

JiTR Demonstration

The JiTR (Just-in-Time Research) prototype is the first pass at realizing the Mashing Texts project. The simple story about JiTR is that it lets you manage collections of digital items and run tools that either gather items, clean them or analyze them. In JiTR you can create collections, add items to collections (manually or with spiders/scrapers), edit the items (automatically or manually) and then pass a collection as a single "text" to a text analysis tool elsewhere. We will demonstrate a rapid research cycle of gathering, editing and analyzing a collection.

Mashing Texts embraces principles of Internet mashups in conceptualizing a recombinant environment whereby the resultant product is more valuable than the sum of its parts. JiTR is designed so that other tools can be plugged into it like search engines and spiders to generate text collections on demand from within the environment. As part of the project's emphasis on malleability of data, JiTR allows those collections to be organized by various tags and labels. The final part of the demonstration will show some of the analysis tools that JiTR offers.

Persona-Scenario Methodology

JiTR is just a working prototype designed to test a design hypothesis about a type of tool that we think is needed. The research in this project has been the extensive design process. If the end purpose is user functionality, why not prototype around stories about users and what they could do? This is the idea behind the personas and scenarios design process, a usability design process that we have used on other projects like the TAPoR portal. Personas are imagined possible users that "act as stand-ins for real users" [3]. Personas are examples of people who would potential like to use the system. They are created to stimulate thinking about people, rather than exclusively concepts. Once realistic user personas are created, usages scenarios are described, which consider ways users could possibly employ the final product. Scenarios move into specifics, detailing the steps that the user would follow in working with the system. Eventually, the various scenarios are prioritized into primary and secondary uses so you know what types of tasks the product needs to support. You can also use the scenarios to audit the prototype.

In this project we started with two constituencies that we came to call DEEP (Distributed Electronic Editing Platform) and BROAD (which does not yet stand for anything) .

- DEEP personas would use JiTR to collaboratively edit a rich born-digital collection like Orlando. To this end we worked with the Orlando team to check that the personas, scenarios, and tasks we described were true to their experience.

- BROAD personas would use JiTR to rapidly gather evidence from the web to study contemporary issues. We imagined various users who want to use the web as their text and therefore want to gather subsets of web documents into study collections. For example, a linguist might want to gather web pages where real users use a language pattern.

Our hypothesis is that both these very different constituencies actually need the same sort of tool and we wanted to see if we could design something that would serve both, another form of mashing, if you will.

This paper will outline the steps of this design process as we believe it is particularly suited to humanities computing. We will end by presenting the priority personas and what these imagined people want to do.

Technical and Political Implications

It is one thing to prototype and test an idea for another tool; it is another to develop a production tool that others can use. We are particularly conscious that a tool like JiTR needs to work within the ecology of tools available. To that end we had a parallel process in the project to identify the other tools, frameworks, and standards that JiTR needs to work with so as to develop viable architectural specification for the development of a production version. This involved identifying the points of articulation between JiTR and other tools. An obvious example is how JiTR should work with repository systems life Fedora. While our prototype has its own MySQL database, a production version should not manage the repository of texts in a collection. Instead it should have the ability to push and pull texts from a repository, whether it be Fedora or another system. Likewise JiTR should not include any tools, but should have a plug-in architecture for tools from spiders to text mining tools. Our prototype has tools built in, but a production system would have an interface for managing processes.

As with any development project the design process is partly about deciding what you aren't going to do. We believe that a missing layer needed between repository tools and text analysis tools is a collaborative research collections management environment. At the end of the paper we will present the designs of how we think a full system could support the research work flow of our two user constituencies. How would an editor develop a workflow for the editing of electronic texts in JiTR? How would a researcher interested in the discourse on the web about high performance computing gather documents, clean them and analyze them?

Notes

[1] Summit on Digital Tools for the Humanities, http://www.iath.virginia.edu/dtsummit/

[2] Mashing Texts is supported by a Social Science and Humanities Research Council of Canada Research and Development Initiative grant. The project is openly documented at http://tada.mcmaster.ca/Main/MashTexts

[3] Calabria, Tina. "An introduction to personas and how to create them", http://www.steptwo.com.au/papers/kmc_personas/

Laying the conceptual foundations for data integration in the humanities

Michele Pasin
Kings College London
michele.pasin@kcl.ac.uk

Arianna Ciula
Kings College London
arianna.ciula@kcl.ac.uk

The purpose of this paper is to promote the discussion on what are the key dimensions of humanities' scholarship, and how they can be best represented by means of formal languages in the context of the Semantic Web. Quite often, available formalizations of knowledge domains and practices in the humanities have been inspired by previous work on more rigorous scientific domains. As a result, we believe that the models thus created tend to oversimplify, if not totally misunderstand, the complexity and peculiarity of the work of humanities' scholars. In this paper, we want to highlight a number of characteristics that need to be taken into account when modeling humanities' data. We argue that only by keeping in mind such requirements we will be able to lay out solid foundations for facilitating non-trivial information integration in humanities domains. We are currently testing these ideas in our department by reflecting upon anumber of preexisting digital humanities projects. The final paper will give a more extensive description of this evaluation.

Introduction

In recent years we have seen a proliferation of research and commercial projects aiming at the dissemination of a large number of structured or semi-structured data. On the academic side, for example, enterprises such as the Semantic Web (Berners-Lee et al., 2001) have long attempted to support the creation of a vast-scale layer of machine-processable data, which should work as an 'extension' of the traditional web. Less academic examples are instead Freebase (Freebase, 2007), a web application aiming at becoming an "open, shared database of the world's knowledge" which can be freely edited by registered users, and the DBpedia (Auer et al., 2007), a community effort to extract structured information from Wikipedia and make it available on the web by means of a public API[1].

In this paper, we associate these developments in web technologies with the term *'semantic web'* (SW), as they all share the intent to encode formally (with varying degrees of complexity) aspects of the meaning of the resources or artifacts they refer to.

It is worth asking then, why should we as digital humanists be adopting a semantic web approach? A primary advantage of having structured data exposed on the web is the possibility to integrate and reuse them in novel ways. For example, we can imagine a scenario where data coming from an archeological project about Tutankhamun are being accessed by other archeologists interested in pottery produced in Egypt in the same period. Pushing it a little further, we could also think of a research group in sociology of science examining the same data, looking for anomalous patterns in the archeologists' daily data-collection practices.

From an examination of the most recent literature, it is easy to conclude that semantic web technologies have already been tested in a variety of domains. These include both hard science domains, such as physics (Friedland and Allen, 2004), biology (Bechhofer et al., 2006), mathematics (Habel and Magnan, 2007), but also humanities' disciplines such as history of art (Hildebrand et al., 2006), literature (Nowviskie, 2005), music (Schraefel et al., 2005).

However, this spectrum of experimentations leads us to a further consideration. Since scientific domains are highly structured they can more easily be mapped into formal conceptual schemas, so as to be used in SW applications -e.g., a *gene* ontology, or an ontology of *hardware components*. This is not the case for all humanities domains, especially where scholars give high value to processes like the expression of *subjective interpretations* and the *debate* on the subject in question, rather than aiming to search for *objective schemas* or *universal taxonomies*. In other words, the task of modeling knowledge domains in the humanities through formal languages (so as to allow computability and data integration) presents various challenges which are still to be tackled by existing research on the semantic web front.[2]

For example, it is our view that systems such as /facet (Hildebrand et al., 2006) or CultureSampo (Eero Hyvönen et al., 2007), although providing advanced interfaces for exploring humanities' data, do not investigate enough the type of semantic 'services' humanities' scholars often engage with in their research practices. In fact, very often such systems make use of very 'shallow' semantic models (e.g., a 'person' who created a 'work' which *belongs-to* a 'style'), thus oversimplifying the actual discourse that makes a statement valuable within a humanities discipline. As a consequence, data thus structured can hardly be of use to the humanities scholar in her research and activities.

If data sharing and integration in the humanities is recognized to be worth pursuing, it is therefore necessary to build some solid foundations for a truly useful semantic web framework in the humanities. The first activity that will help us in this respect is a thorough consideration of the typical *entities* and *practices* emerging in humanities' research. Accordingly, in section 3 we outline a number of key requirements humanities' semantic models should support. In the following section we spend some words on the approach that drivesour usage of ontologies for data integration.

Ontology: a beauty or a beast?

A central notion in the semantic web and in the world of data integration is that one of *ontology*. The widely used definition by Gruber (Gruber, 1993) describes it as "an explicit specification of a conceptualization". Being a *conceptualization* an ontology is therefore a *stylized representation* of the world; secondly, since it is expressed in a *formal* language, an ontology can be defined *unambiguously*. As a consequence, ontologies are well suited representation languages for describing data and sharing information; their employment is also endorsed by the W3C (W3C, 2004).

Besides this quite conventional view of what an ontology is, the debate is ongoing about the status of an ontology with respect to the world it represents. For example, some authors such as Smith (Smith, 2003) hold a *realist* position, while others such as the aforementioned Gruber (Gruber, 2003) support a more *pragmatic* view. Such positions affect inevitably the way ontologies are developed and used. For example, in the first case (realist) the implicit assumption is that the ontology should approximate to a 'true' reality; as a consequence, multiple ontologies about the same subject should ultimately converge in their modeling choices. On the contrary, the second class of ontology-design approaches (pragmatist) seean ontology essentially as an engineering artifact: thus, it does not hold any absolute value about the reality it depicts, but it provides a practical solution to the 'problems' it was designed to tackle (i.e. it is a *mean* to an *end*).

Although in the SW world both approaches have many followers, the context in which digital humanities practitioners and researchers operate, in our opinion, is much closer to the pragmatic approach. Indeed, the humanities are often perceived as the place where all the voices —provided they are respectful of certain argumentative conventions—can be heard, and where all the assumptions can be questioned. Therefore, ontologies for the humanities must support *diversity* and *variety* of *viewpoints*; thus they cannot adhere to an underlying model which neglects multiplicity in favor of a monolithic vision of the world.

Following Gruber (Gruber, 2003), we therefore intend to promote the concept of an ontology as the *agreement* reached by multiple *parties* (e.g., programmers, scientists, collaborators, librarians) with the aim of accomplishing some *objectives* (e.g., data exchange between applications, communication between people, integration of disparate representations). Using a metaphor, ontologies are *contracts*, they are the *currency* used to perform some valuable operations. Thus, their importance is ultimately related not to their truth or beauty, but to the ease they bring to the collaboration among people[3]. To use a less 'commercial' metaphor an ontology is a compromise or a point of contact between specific and possibly divergent models. The issue is therefore not only to identify commonalities between projects, for instance, but also to agree that the compromises so found won't diminish the value of the underlying idiosyncratic models, the specificity of any single project or interpretation. We believe that in the humanities this agreement is not necessarily reachable once for all or hoped for, because it may imply the negation of the interpretative efforts that make a work or a project unique and the negation of the evolutionary nature of scholarship. However, we also think that the possibility to make two incommensurable categorical systems communicate could be a challenge worth pursuing.

Defining humanities' research

As mentioned above, at a general level it is useful to characterize humanities scholarship by highlighting the points of contrast with the hard sciences. Humanities scholars are traditionally engaged with the expression of interpretative statements and the elaboration of debates on a disparate range of sources of knowledge, rather than with the seeking of firm objective schemas or universal taxonomies. One of the authors analyzed more specifically the characteristics of a humanities domain (Pasin et al., 2007)—philosophy—and identified those key elements that define its scholarship and make it hard to model. Some of these elements are outlined below:

1. **historical** events, that is, events which are inherently time-dependent (e.g., the publication of a book, or an author's subscription to a viewpoint);

2. generic **uncertainty**, that is the frequency of statements about facts which cannot be located exactly in the time and space dimensions (e.g., the birth of Heraclitus);

3. **information objects**, i.e. *texts* in a semiotic sense and especially language-based information objects (e.g. a book), as they are the traditionally preferred medium philosophical contents are expressed with;

4. **interpretation** events, intended as the process of attributing an abstract content to an information object (e.g., when we say that "Aristotle's fourth book of the Metaphysics states an ontological principle");

5. coexistence of **contradictory** information, which is a direct consequence of 4 (e.g., when people claim different or opposing views on the same proposition);

6. **viewpoints**, and other non-material entities ("philosophical ideas"), for they are the objects philosophers are usually engaged with by studying and expressing them.

Although philosophy has often been defined as the queen of the sciences, these reflections on its nature as discipline may not stand true for thehumanities as a whole. In order to highlight all the dimensions that make the modeling of humanities domains such a unique task, we surely need a thorough investigation of other humanities' domains too. Furthermore, for space reasons we have deliberately not mentioned other works in the digital humanities, such as (Jones, 2006) and (Eide, 2008), where the issues tackled are remarkably similar to ours, although the approach is not necessarily ontology-oriented. We intend to elaborate more on these topics in the final version of the paper.

Conclusions

In this extended abstract we addressed a number of problems emerging from the employment of semantic web technologies in humanities domains. In particular, we focused on the notion of *ontologies for data-integration*, highlighting the great challenges these technologies will bring *especially* to the digital humanities' practitioner. To this aim we also provided some examples from our previous research in the philosophical domain. In the final paper we will expand this research also by drawing from the results of a detailed analysis of the various projects ongoing in our department. It is our hope that this research will stimulate further discussions and the formulation of a preliminary but comprehensive research agenda.

References

Auer, S. et al. Dbpedia: A Nucleus for a Web of Open Data. *6th International Semantic Web Conference* (ISWC 2007) (2007).

Bechhofer, S., Stevens, R. D. & Lord, P. W. Gohse: Ontology Driven Linking of Biology Resources. *Web Semantics: Science, Services and Agents on the World Wide Web* **4**, (2006).

Berners-Lee, T., Hendler, J. & Lassila, O. The Semantic Web. *Scientific American* (2001).

Eero Hyv√∂nen et al., CultureSampo-Finnish Culture on the Semantic Web. The vision and first results, in *Information Technology for the Virtual Museum* (ed. Robering, K.) (LIT Verlag, 2007).

Eide, √ò. The Exhibition Problem. A Real-Life Example With a Suggested Solution. *Literary and Linguistic Computing* 23, 27-37 (2008).

Freebase. An open, shared database of the world's knowledge. (2007), Retrieved 20 Feb. 2009, http://www.freebase.com/

Friedland, N. S. & Allen, P. G. Project Halo: Towards a Digital Aristotle, (2004). Retrieved 20 Feb. 2009, http://www.projecthalo.com

Gruber, T. It Is What It Does: The Pragmatics of Ontology. *Invited presentation to the meeting of the CIDOC Conceptual Reference Model committee* (2003). http://tomgruber.org/writing/cidoc-ontology.htm

Gruber, T. R. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* **5**, 199-220 (1993).

Habel, G. & Magnan, F. General Poncelet Meets the Semantic Web: A Concrete Example of the Usage of Ontologies to Support Creation and Dissemination of Elearning Contents. *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007* 908-915 (2007).

Hildebrand, M., van Ossenbruggen, J. & Hardman, L. /Facet: A Browser for Heterogeneous Semantic Web Repositories. *International Semantic Web Conference* (ISWC2006) (2006).

Jones, A. (ed) *Summit on Digital Tools for the Humanities: Report on Summit Accomplishments*. (2006). Retrieved 20 Feb. 2009, http://www.iath.virginia.edu/dtsummit/SummitText.pdf

Nowviskie, B. COLLEX: semantic collections & exhibits for the remixable web. (2005). Retrieved 20 Feb. 2009, http://www.nines.org/about/Nowviskie-Collex.pdf

Pasin, M., Motta, E. & Zdrahal, Z. Capturing Knowledge About Philosophy. *Knowledge Capture* (KCAP) (2007).

Schraefel, m. c. et al. The Mspace Classical Music Explorer: Improving Access to Classical Music for Real People. *V MUSICNETWORK OPEN WORKSHOP: Integration of Music in Multimedia Applications* (2005).

Smith, B. *Ontology*, in Blackwell Guide to the Philosophy of Computing and Information (ed Floridi, L.) (Blackwell, Oxford, 2003).

W3C. OWL Web Ontology Language Overview. (2004). Retrieved 20 Feb. 2009, http://www.w3.org/TR/owl-features/

Notes

[1] Application Programming Interface, that is, an access point by which such data can be retrieved or manipulated programmatically.

[2] It has to be noted that some of these challenges have been faced in previous efforts (preceding the advent of the web) of formalizations in the humanities: good examples are the creation of domain-specific thesauri and taxonomies, or the classification systems in library studies.

[3] Note that we are focusing on the conceptual implications here rather than on thechallenges of an ontology implementation by using specific computer languages.

The Limit of Representation

Elena Pierazzo
King's College London
elena.pierazzo@kcl.ac.uk

In the proposed paper, I reflect on the theoretical implications and methodology of diplomatic digital editions which have arisen out of a three-year AHRC-funded project devoted to Jane Austen's holographic fictional manuscripts. The title of the paper is deliberately ambiguous in order to mark the dual nature of a topic that applies to both transcription and publication.

For transcription the *TEI Guidelines* offer a large variety of elements to transcribe and describe primary source documents at almost any level of granularity and detail: "to all intents and purposes", Driscoll remarks, "there is no limit to the information one can add to a text – apart, that is, from the limits of the imagination" (2006:261). But the availability of an element does not imply the necessity of using it—and this is a crucial point, as we will see[1].

For publication, Tanselle notes, "the editor's goal is to reproduce in print as many of the characteristics of the document as he can" (1978: 51). Although this may be valid for print, it is of little help in a digital environment, where you can represent much more[2]. This can be positive, e.g. to avoid most of what Michael Hunter calls the "confusion of brackets"[3], but it leaves open the question of where to stop – a question only made more complicated by the extensibility of the TEI.

Driscoll states that:

> In the determination of how much information should be included [in the transcription of a manuscript], the decisions facing the producer of an electronic transcription are essentially the same as those facing the producer of any transcription (2006:257-8).

However, even if the decisions are of the same kind in both media, they may have quite different resolutions for a digital edition, as the new medium allows the transcription and output of many more features, as I just noted.[4] Nevertheless when designing a digital transcription a scholar needs to define her/his own boundaries:

> An electronic edition is like an iceberg, with far more data potentially available than is actually visible on the screen, and this is at the same time a great opportunity and a temptation to overdo things. When so many possibilities exist, there is a danger of technological considerations of what can be done taking priority over intellectual considerations of what is actually desirable or necessary in any particular case. (Hunter 2007:71).

But when does stopping imply a loss of information and when is it simply wise? More to the point, what is the information that needs to be represented and which is desirable but not essential? Can this essential information be feasibly encoded in the transcription? And can it be conveniently represented in the visual output? And again, if the edition includes a digital image of the source, should the limits change or is the diplomatic edition altogether worthless[5]?

The answers to most of the previous questions depend on the research goals and the editor better judgement in interpreting which feature contained in a source is relevant, and which is merely ornamental: as Claus Huitfeldt demonstrates, in fact, there is not such a thing as an objective diplomatic transcription (2006:194-6). Nevertheless those involved with print have been able to produce guidelines and lively discussions, while their counterparts in the digital medium seem more or less to be avoiding the topic[6].

According to Tanselle, essential features that need to be retained in a diplomatic transcription of modern holographic documents are: punctuation, spelling, letter shape (long s, i-j, u-v, for instance), capitalization, abbreviations, authorial errors, deleted readings, added texts (maintaining or marking their positioning). In a digital framework we can do more than that, for instance we can measure dimensions of gaps, distinguish a wider range of letter shapes, reproduce the colour of the ink, the temporal sequence of the revisions, etc. On the other side Hunter does not hide his uneasiness toward typefacsimile editions, stating that: "the aim is to produce an edition which does justice to the content of the manuscript, paying attention to its actual appearance but not fetishising this", adding, though, that as markup languages allow the editor "to have his or her cake and eat it", it is possible to push the representation forward and, for instance, "if the editor sees a value in preserving the abbreviated forms […] they are welcome to do so" remarking though that "I would not bother" (2007:85, 80).

The XML markup language, especially when used according to the TEI Guidelines, allows editors to encode features that could serve for different displays of the text. Thus from the same text we can easily produce diplomatic, semi-diplomatic, reading and edited texts. We can have our cake and eat it. But encoding for multiple outputs is sometimes not very easy; compromises may be required. In the transcription of Jane Austen manuscripts we see, for instance, interlinear insertions being placed in positions that do not correspond to logical insertion points: in this case we had to chose which aspect to privilege, the semantic of the text or the appearance of the source document, the former being relevant to the production of a reading edition and the latter for the diplomatic edition.

It is clear that there are limits to the representation of documents, and those limits are both conceptual and pragmatic. Once the document has been transcribed, a certain level of distance between the physical object and the transcribed object is inevitable. Hunter again:

> different handwritings and letter forms, but also ink blots, different methods of striking through words, or exact details of layout, for which only a pictorial facsimile suffices. The chief thing which a type-facsimile can do is to distinguish words in pen or pencil, or in different hands, but even this might be better achieved by a commentary on a photographic or digital reproduction (Hunter 2007:75).[7]

In fact, no transcription, however accurate, will ever be able to represent entirely the source document. Some characteristics of the manuscripts are irremediably lost by transcribing it, e.g. the variable shape and spacing of handwritten glyphs versus the constant shape of digital fonts. As Hans Walter Gabler says, "clearly the diplomatic transcription is already a distinct abstraction from the document" (2007:204). On the other hand, the more details we add to our transcription and the more accurate it is, the greater the density of the markup, with consequent loss of readability and loss of editorial control over the text. Even if stand-off markup can help to address this problem, the "cost" of the markup remains relevant.

This paper will use the experience of the Jane Austen Project to address the following questions from a theoretical and pragmatic point of view:

- What cannot be represented in a transcription?
- What can be represented but at too high a price (i.e. requiring too high a level of encoding) to be feasible?
- What should not be represented?
- Which features can be encoded but cannot be reproduced with existing digital publication standards (i.e. HTML/CSS)?
- To which display should a given feature belong?

The paper will open up discussion on essential aspects of editing texts in a digital environment by attempting to define a minimal shared protocol in digital transcription of documents. The author hopes that many other contributions will follow.

References

Driscoll, Mattew J. "Levels of Transcription". In *Electronic Textual Editing*. L. Burnard, K. O'Brien O'Keeffe and J. Unsworth (eds). New York, 2006, pp. 254-61.

Gabler, Hans Walter. "The Primacy of the Document in Editing". *Ecdotica* 4 (2007), 197-207.

Hunter, Michael. *Editing Early Modern Texts: an introduction to principles and practice*. New York, 2007.

Huitfeldt, Claus. "Philosophy Case Study". In *Electronic Textual Editing*. L. Burnard, K. O'Brien O'Keeffe and J. Unsworth (eds). New York, 2006, pp. 181-96

Kiernan, Kevin. "Digital Facsimile in Editing". L. Burnard, K. O'Brien O'Keeffe and J. Unsworth (eds). New York, 2006, 262-8.

Robinson, Peter. "Where we are with Electronic Scholarly Editions, and where we want to be". *Computerphilologie* 4 (2002); available at http://computerphilologie.tu-darmstadt.de/jg03/robinson.html.

Sperberg-McQueen, C. Michael. "Text in Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts". *Literary Linguistic Computing* 6 (1991), 34-46.

Tanselle, G. Thomas. "The Editing of Historical Documents." *Studies in Bibliography* 31 (1978): 1-56.

Notes

[1] Sperberg-McQueen 1991:36: "When I observe […] that X and Y must be tagged (or taggable), I mean not that everyone should be required to tag X and Y, but that any general-purpose markup scheme must provide tags to enable researchers to tag X and Y when they wish so".

[2] And indeed Robinson 2002 remarks his frustration toward digital editions that do not present "material in a manner significantly different from that which could have been managed in print".

[3] For instance, speaking of square brackets, he says: "these have been used for almost diametrically opposite purposes – to indicate deletions in the original; to denote text lost through mutilations; or to denote editorial sup-

ply." See Hunter 2007:118-20

[4] See, for instance the so-called "ultra diplomatic transcription" which integrates facsimile images and transcription, used by the HyperNietsche Project; an example of which is to be found in Gabler 2007:205 at http://www.hypernietzsche.org/demo/bksailehwgabler-33.

[5] "The image-based scholarly edition subsumes the purpose of a diplomatic edition and removed the fruitless frustration of trying to preserve the exact layout, illumination, and physical appearance of a manuscript in printed form. "(Kiernan 2006:266)

[6] But see Huitfeldt 2006:190.

[7] See also Sperberg-McQueen 1991:34: "What computer process are representations of data. […] Representations are inevitably partial, never disinterested".

The Atlas of Early Printing: Digital History and Book History

Gregory J. Prickman
The University of Iowa Libraries
greg-prickman@uiowa.edu

The history of the book is an interdisciplinary field of inquiry that has been emerging for the last decades. Scholars of the history of the book use the development of books and reading to study larger historical ideas. In the digital realm, the digitization of books has received a great deal of attention, including funding and media coverage, as facsimiles of physical books are created for the digital environment. The application of digital tools to the broader study of the history of the book, however, has been much slower to catch on. This presentation will introduce the *Atlas of Early Printing*, a digital project utilizing tools such as GIS data, Flash-driven interactivity, and 3D computer graphics visualization to transform how information about the invention and spread of printing during the fifteenth century is presented. Rather than a consideration of the digital book, the *Atlas of Early Printing* represents new directions in the digital book history.

The background of the project will be outlined by an overview of the historical moment the *Atlas* is designed to represent. The printing press was developed by Johannes Gutenberg during the mid-fifteenth century in Mainz, Germany. Following the production of his famous Bible during the years 1450-1455, printing slowly began to spread throughout Europe as the secrets of the trade were handed down to apprentice and journeymen printers who set off to establish businesses. By the beginning of the sixteenth century, printing was a well-established and widely accepted trade, with presses operating in hundreds of European towns and a brisk trade in books increasing in volume every year.

The spread of printing has been depicted in several maps, the most well-known being those in Lucien Febvre and Henri-Jean Martin's *L'apparition du livre*. In these views, the arrival of printing in towns across the continent is arranged by decade, showing the broad trend of movement but lacking the detail necessary for more sophisticated interpretation. In addition, the maps lack any contextual information. For an event as revolutionary as the invention and adoption of printing, this context is crucial to an understanding of the forces at work: what elements in society supported printing and ensured its success? What access did printers have to resources and markets? How did these complicated factors work together to produce the intellectual environment for printed books to flourish?

The *Atlas of Early Printing* offers an interactive map that not only animates the spread of printing year-by-year, but includes layers that place printing within a historical and cultural context, such as the locations of paper mills, universities, market towns, and trade routes. All of these layers can be controlled by users, allowing them to view as much or as little information as they choose. The site relies on Flash to display the information online, drawing from a series of XML files containing the data. Changes to the site can be made instantaneously by changing the data in the XML files, leading to a flexible and scalable site that does not require extensive database maintenance. New layers can be added by creating a new XML file, without disturbing the information already present.

The presentation will examine these technical aspects of the map before shifting focus to another primary feature of the site: a 3D computer graphics model of an early printing press (http://atlas.lib.uiowa.edu/press-animation.html). The model was created in Maya and can be rendered in any number of views and animations. The technical process of creating a printed page from metal type inked and pressed on a wooden structure is essentially foreign to many students and even scholars of history. While physical replicas can be viewed at several museums and libraries around the world, few people have the opportunity to experience their operation in person. The digital model brings this esoteric history to life in a manner that enables a user to see a press in action. Because the press is modeled in Maya, individual pieces can be modified. Any reconstruction of an early printing press is conjecture, so the flexibility of the 3D model allows for changes and variations. Future plans include detailed close-up views and, ultimately, a model that a user can manipulate in real time online.

This presentation will also describe the site's development process. The primary goal for the project was to create an intuitive, easy-to-use, yet in-depth resource with widely available software in a compressed period of time. The *Atlas of Early Printing* was created over the course of a year from initial discussion and funding to the uploading of version 1.0. The collaborative effort between different university units that had seldom worked together previously, which enabled this schedule, will be discussed.

Finally, the possibilities for future expansion and development will be considered in light of the potential that digital mapping holds for a broader range of book history topics. By applying digital humanities tools and techniques to the databases and image resources representing historical books, new ways of capturing and depicting the history of books as objects can be achieved. The *Atlas of Early Printing* is an initial step toward projects that will analyze bibliographic data and combine it with geographic information systems, to provide methods for researching and teaching the history of the book.

Embedded Text Analysis

Brian L. Pytlik Zillig
University of Nebraska-Lincoln
bzillig1@unl.edu

John Unsworth, in writing about the rhetorical model in the humanities, asserts that "we believe that by paying attention to an object of interest, we can explore it, find new dimensions within it, notice things about it that have never been noticed before, and increase its value." [1] One way that digital humanists pay attention to texts is to assemble them into groups and analyze them. In the last few years, the digital humanities have increasingly benefited from the ability to perform analytical operations on individual texts and, less frequently, on text corpora. Efforts such as TaPOR, HyperPo, Nora Project, WordHoard, TokenX, and others have made it possible, with varying degrees of success, to process texts and output a wide assortment of data about them. The MONK Project has worked to combine full-text archives so that the once-exclusively offline activity of performing sophisticated algorithmic analyses on large text corpora can be performed, in a web browser, for large sets of documents.

Outside of the digital humanities realm, some commercial sites are beginning to offer access to limited text data. In 2006, the New York Times, using technology developed by Answers.com, quietly implemented the JavaScript "double-click dictionary" feature, where users could double-click any word in an online article and a new window would pop up with dictionary and thesaurus information [2]. While this may not serve as a traditional example of text analysis per se, it moves closer to new dimensions of noticing.

These efforts, both within digital humanities and without, have gone a long way toward making text analysis techniques accessible to a somewhat wider audience. But, as the maintainers of most of these projects would probably acknowledge, this task is made more difficult by the complexities associated with, and diversity of, the analytical techniques desired. Developers in this area have traditionally employed a specialized vocabulary, designed interfaces that are anything but general, and placed a high premium on information density in their representations. This ideal is not always in accord with the design of an interface intended for reading a text. In fact, there are significant problems associated with presenting text analysis to users. In one example, the interface I developed for TokenX suffers from the same shortcomings that many web-based text analysis applications have: (1) it is insufficiently intuitive, (2) users have to read the menu/link options because there is no universally acknowledged visual language of text analysis, (3) there are often too many steps involved in getting from where you are to where you wish to be, and (4) the user must summon a new page, or parts thereof, to see the results of a given text analysis operation. [3]

There is a clear need for nearby points of entry to text analytical results—ones that are conveniently embedded in the document the user is reading. I use the term "nearby" in contrast to "convenient" to emphasize the collocation of texts and textual data. It is not sufficient that data be a few clicks and scrolls away from where you are when you are reading. The data must be where you can see it at the point you decide you want it. An example of such a point of entry that is sufficiently proximal might be the browser tooltip. A conventional use of a tooltip involves an information-bearing mouseover effect that appears on request and disappears when the mouse moves away. To a similar end, I propose the creation of a text analysis interface that may seem radical. Such an interface would require no configuration, avoid rarefied terms and procedures, and stand in an immediate relationship to the reading field of the text under investigation. It would be located where the user is reading at a given moment. From the perspective of the user, the analysis would be located in the document, or corpus, itself. No longer would it be necessary to travel to a different online site to generate text data about either the text that the user is reading or the corpus it occupies.

Embedded text analysis would, for the sake of broad interoperability and access, function in a wide variety of modern browsers and would afford users access to quantitative data about individual documents and, where appropriate, the corpus of texts related to the present document. Unlike other text analysis tools, an embedded text analysis interface will not sever the connection between the reading text and the data, but will foreground it. For reasons presumably similar to those underlying the embedding of the electric washing machine into the average home in the 20th century, embedded text analysis will emphasize the convenience of the user. From the perspective of the user, embedded text analysis features should seem clear and obvious.

Technical challenges abound, ranging from how best to embed complex data results in ways that are easily understood and do not interfere with reading to identifying in advance the analytical procedures that a user might want to invoke. Such procedures would probably include lemma information, part of speech, word n-grams, character n-grams, lemma n-grams, word and character counts, keyword and n-gram in context visualizations, word distribution, term frequency–inverse document frequency (TF-IDF) data, sentence lengths, and more.

An approach to data visualization using embedded text analysis techniques is innovative, both in terms of development and deployment, and anticipates future needs for emerging methods of looking at large sets of text data. As more humanities content is digitized and made available—the entire Text Creation Partnership collection of texts will be freely available in 2015—there will be more demand for tools to perform increasingly sophisticated analyses. The historic manner of representing text data as a static graphic element, rooted in print publishing and its once-necessary reliance upon static representations of data, will yield to a growing need for dynamic user-directed visualizations embedded precisely where they are needed: inside the reading field itself.

References:

1. Unsworth, J. (2006) "Digital Humanities Beyond Representation," University of Central Florida, Orlando, FL November 13, 2006. http://www3.isrl.uiuc.edu/~unsworth/UCF/

2. Answers Corporation (2006) "NYTimes.com Integrates Answers.com Reference Content."

3. TokenX was designed, at the University of Nebraska-Lincoln's Center for Digital Research in the Humanities, to provide an easy-to-use web-based interface for text analysis and visualization that supports both XML and TEI texts.

The Ghost in the Manuscript: Hyperspectral Text Recovery and Segmentation

Patrick Shiel
National University of Ireland
pshiel@cs.nuim.ie

John G. Keating
National University of Ireland
john.keating@nuim.ie

Malte Rehbein
National University of Ireland
malte.rehbein@nuigalway.ie

Introduction

The condition of medieval manuscripts ranges from those that are fully legible to those which can only be read in part, and their legibility is determined by the manner in which they were preserved and treated throughout the ages. In some cases deterioration is due to processes such as fading or staining; in others, the text may have been interfered with in some way. For instance, in the Irish context, the oldest surviving manuscript written entirely in Irish, Leabhar na hUidhre (12th century) was subject to part-erasure and rewriting by a scribe who was active at some point between the 12th and the 14th centuries. The Yellow Book of Lecan (circa 1400) was overwritten in part by an antiquarian scholar active in the 18th century, while the use of chemical agents to enhance readings by scholars working in the 19th century has had a long-term detrimental effect on a number of other manuscripts. Two central research questions of interest to all scholars, therefore, are:

- To what extent is text recovery in key medieval texts possible using current technologies, e.g. palimpsest, fading, deliberate removal?
- To what is it possible to establish irrefutable scientific evidence for interpretation of questioned documents, e.g., identify the different hands (inks)?

In this paper, we illustrate how to provide answers to these questions using modern scientific techniques and emerging forensic technology, i.e. hyperspectral imaging [1] and associated image processing techniques [2].

The acquisition of a hyperspectral scanner by An Foras Feasa (AFF), a Forensic XP-4010, has presented an opportunity to subject damaged or illegible texts to a modern scientific re-examination. The scanner has the potential to read various different layers of a manuscript in a manner not possible to the human eye and to analyse elements of its composition. As such it presents the possibility of retrieving text that has been lost through fading, staining, overwriting or other forms of erasure. In addition, it offers the prospect of distinguishing different ink-types, and furnishing us with details of the manuscript's composition, all of which are refinements, which can be used to answer questions about date and provenance. This process marks a new departure for the study of manuscripts and may provide answer many long-standing questions posed by palaeographers and by scholars in a variety of disciplines. Furthermore, through text retrieval, it holds out the prospect of adding considerably to the existing corpus of texts and to providing very many new research opportunities for coming generations of scholars. In this introductory paper on hyperspectral imaging we concentrate on two key processes: text recovery and text segmentation.

Methodology

The investigative and analytic methods described here are based on a novel and highly specialised technique called Hyperspectral Imaging. Hyperspectral imaging is a non-destructive optical technique that measures reflectance (fraction of light reflected) characteristics of a document with high spatial and spectral resolution. An hyperspectral imaging device records a sequence of digital images of the selected manuscript area (with maximum dimensions 50 mm x 50 mm) illuminated with monochromatic light from a tuneable light source from 350 nm (near-UV) through the entire visible range and up to 2400 nm (infrared). The value of each image pixel in the recorded image sequence represents an accurate measurement of the reflectance curve for a tiny, 13 micron square, area on the document. Analysis of all spectral curves, essentially a cube of information, provides information about the physical characteristics of questioned manuscripts.

Hyperspectral imaging, together with modern two-dimensional spectrum software and three-dimensional image and visualisation software, provides modern researchers working in the field of historic documents analysis with opportunities for forensic examination that were heretofore unavailable. Methodologically, there are two main fields of applications of this technique: (i) the extraction of relevant historic, diplomatic and palaeographic information from documents and (ii) the investigation of the impact of environmental conditions on document condition and of degradation effects on writing materials and substrates. In particular, reflectance curves found in different sections of the manuscripts can be compared with each other in order to determine whether different types of inks had been used during text composition or to identify modifications that occurred during the manuscripts' history. Light spectroscopy analyses may also be conducted to aid recovery and segmentation. Fluorescence occurs when an object emits a high wavelength (low energy light) following illumination by a shorter wavelength (higher energy light) due to molecular absorption of part of the incident light. Furthermore, the spectral curves may be compared with those in international databases containing typical ink spectra to determine and date the kind of ink or pigment used. The image cube recorded using the technique may be used to enhance the visibility of hidden material such as palimpsest or erased text.

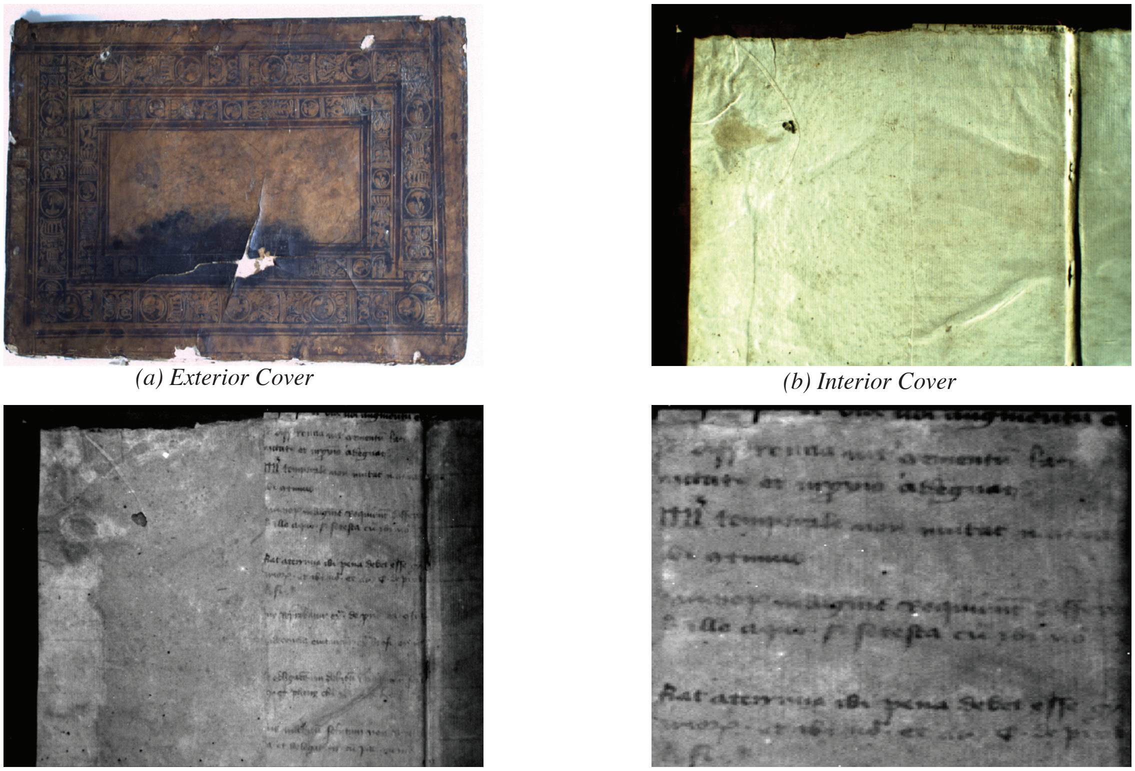
This methodology for manuscript analysis is of significant interest to archivists. For example, The National Archive of the Netherlands in The Hague recently commissioned the development of a dedicated hyperspectral imaging instrument specifically designed to analyze archival documents and to monitor their aging process during storage and exhibitions. Furthermore, at the 16th ICA (International Congress on Archives) Congress in Kuala Lumpur from 21–27 July 2008, two special sessions were devoted to the application of hyperspectral imaging for historic document analysis and its use in providing quantitative, objective criteria for balancing the competing interests of document preservation and presentation. This proposal is therefore informed by current international research and best practice in document analysis and textual imaging.

Example: Text Recovery of hidden text

Figure 1(a) shows a 16C book cover that has become degraded with time and contains mould in places. The interior cover's structure, shown in Figure 1(b) consists of an underlying text which has been pasted over with a clean blank faced sheet of paper. Using fluorescence spectroscopy, i.e, light induced fluorescence in the pages it is possible to reveal the underlying text as shown in Figure 1(c) and 1(d). Here, we will describe the underlying fluorescence and reflectance spectroscopy principles and techniques used to reveal hidden text of this nature, and in several other examples. A complete description of the techniques used for this particular recovery is given in another paper in this conference session.

Example: Text Segmentation of three different inks

The general objectives of cursive text segmentation in-



(a) Exterior Cover

(b) Interior Cover

(c) Recovered Text

(d) Enlarged Recovered Text

Figure 1. Hyperspectral recovery of hidden text fom pasteboard down

clude tasks such as word spotting, text/image alignment, authentication and extraction of specific fields [3]. An important step associate with all of these tasks document segmentation into text lines. In general, this is diffcult due of the low quality and the complexity of these documents, and automatic text line segmentation is an open research field. In general, sophisticated image processing of single-image documents is the norm [3].Here we describe recent approach to the segmentation problem, which we refer to as *hyperspectral segmentation*; the technique is based on the separation and segmentation of differing inks by recording and analysing the reflectance properties of different inks. This technique is particularly useful for the segmentation of texts that have been edited by various authors over a long time period. It is also possible to date the editions by comparing the segments with known dates, or using repositories containing hyperspectral properties of different materials (e.g. inks, paper, etc.). Figure 2, below, shows the results of the technique when used to segment some text with two later annotations (giving three primary additions or segments). Three different inks are used, although the all appear similar in visible light, as shown in Figure 2(a). Figure 2(b) shows the reflectance curves for the two annotations and how it is possible, using appropriate image processing to colour-code the different inks. Figure 2(c)demonstrates how it is possible to foreground the original text, and Figure 2(d) demonstrates that is is possible to isolate, or segment, the annotations. A complete description of the potential for these hyperspectral segmentation techniques for dating and segmenting the *Göttingen kundige bok 2* is given in another paper in this conference session.

Conclusion

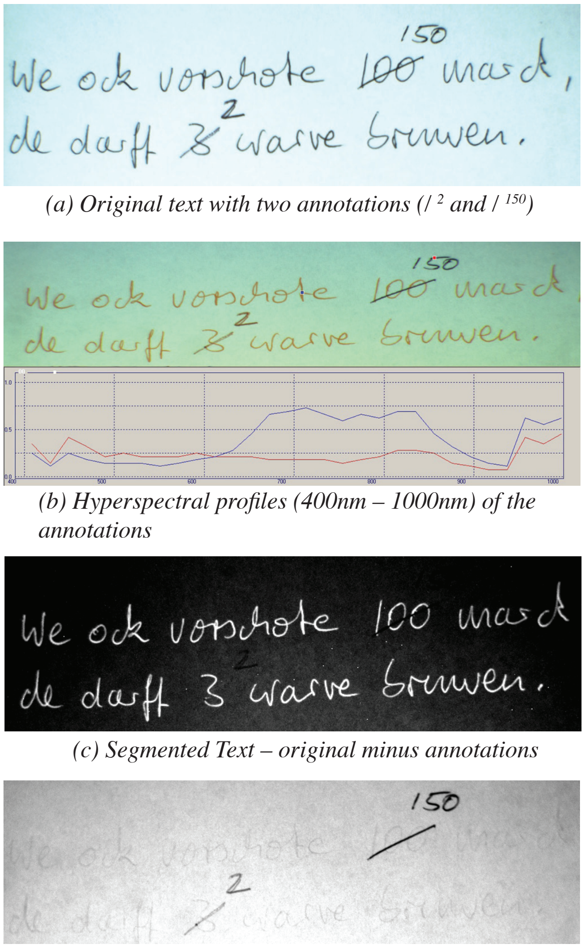
The recovery of illegible or deleted text, the analysis and dating of ink-types, and the minute analyses of scribal hands, segmentation of cursive test, are activities in palaeographic and manuscript studies, and their subsequent textual encoding. This abstract describes how hyperspectral imaging can be used to perform quality text recovery, segmentation and dating of historical documents. Our paper will provide a complete overview of the acquisition (reflectance spectroscopy, fluorescence) , computational and segmentation techniques used for (i) a 16C pastedown cover, and (ii) a three-ink example typical of that found in *Göttingen kundige bok 2*.

References

1. Chang, C.I.: Hyperspectral Imaging: Techniques for Spectral Detection and Classification. Kluwer Academic, New York, N. Y. (2003)

2. Chang, C.I.: Hyperspectral Imaging: Signal Processing Algorithm Design and Anal¬¨ysis. John Wiley and Sons, New York, N. Y. (2007)

3. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: a survey. International Journal on Document Analysis and Recognition 9(2-4) (April 2006) 123–138



(a) Original text with two annotations (/ ^{2} and / ^{150})

(b) Hyperspectral profiles (400nm – 1000nm) of the annotations

(c) Segmented Text – original minus annotations

(d) Segmented Text – two visibly different annotations

Figure 2: Hyperspectral segmentation of sample text – original and two different annotation inks

The Electronic Broadway Project

Doug Reside
University of Maryland, College Park
dreside@umd.edu

Many of those whose profession or study is now in the field of musical theater recount a similar childhood memory of listening to a library copy of a cast album while following along with a printed libretto. Although this is an awkward and incomplete way of experiencing the historical works of musical theater, it still is the only way many fans of the genre not lucky enough to live close to New York can read the text of new Broadway musicals. Commercially released filmed versions of stage shows are relatively rare and do not allow for the careful close reading of the texts students of other forms of drama take for granted. The Electronic Broadway Project, based at the Maryland Institute for Technology in the Humanities (MITH) and funded by an NEH digital humanities startup grant, seeks to provide a better way of accessing the important texts of musical theater for the next generation of fans, scholars, and artists who may even now be pausing their MP3 player and rereading a particularly moving lyric.

In this paper, I will describe the technical, legal, and scholarly challenges the project team faced and overcame as we strove to create an electronic edition of the new Broadway musical *Glory Days*. *Glory Days* was, in many ways, an ideal first title for our series as it poses many difficult research questions with implications that extend beyond musical theater studies alone. Like so many new literary works, *Glory Days* was mostly written using digital tools (word processors, digital music recorders, etc) and so the primary sources are, in many cases, preserved as bits on magnetic media rather than as ink on paper. The creators kept meticulous archives of the various versions of their script during rehearsals, and so the work of the critical editor lies in organizing and collating electronic, rather than manuscript drafts (a process which will become increasingly common in the future of editing and textual criticism). The creative team also preserved their instant message and email discussions, thereby creating a record of the creative process theater historians of earlier periods would envy.

Further, the collaborative nature of musical theater creates a situation in which many different parties hold copyrights for the component elements, each of whom must be convinced to allow access to their work for educational purposes. This is a difficult problem, and one perhaps best solved by clear precedents that shift the cultural atmosphere from one of paralyzing fear of piracy to an understanding that open access to the texts of musical theater can promote interest in licensed productions of their works. The copyright holders of *Glory Days* have generously given us permission to produce a free and publicly accessible edition of their work, an act that, we hope, will set such a precedent.

Of course, the biggest technical challenge in producing an electronic edition of a musical is finding an intuitive way to represent and link the various languages (musical, verbal, and terpsichorean) the form uses to communicate. The web-based interface the project team designed will allow the user to have complete control over the amount of information available on the screen at all times. Any number of windows containing various objects from the digital collection can be opened (or closed) as the user requires. Clicking on a lyric or a bar of music will start the associated recording playing from exactly that point in the song. Notes in the text will link to critical commentary, images of artifacts in the archive related to a particular line, or variants in other versions. Actors will be able to choose to display only the lines and cue lines for a particular character in order to learn their lines. Musicians, in future versions if not in this prototype, will be able to play along with a digital orchestra that includes every part but their own.

Although our edition of *Glory Days* will focus on the text and music of the work, we will also consider the difficult problem of representing the visual elements of the piece and design our interface with this sort of extensibility in mind. How, for instance, does one present a scholarly edition of a dance number? Is there an inexpensive way to digitally capture and render motion in order to create an accurate transcription of dance in the same way that text and staves now encode spoken words and music? Because musicals tend to privilege spectacle and are often developed with the limitations and possibilities of a particular theater in mind, the performance space itself should be considered a part of the piece, but how does one highlight this spatial component of a work in an edition (electronic or otherwise)? How should the performance of music and dance best be linked to their coded representation on scores and choreographers notes? In this paper I will explore such questions and describe the project teams best plans for answering them in phase two of our project.

The Shakespeare Quartos Archive and TEI-P5

Doug Reside
University of Maryland, College Park
dreside@umd.edu

The Shakespeare Quartos Archive will make available every extant copy of the quartos printed prior to 1641, starting with Hamlet. We will further provide transcripts, marked in TEI P5, of each of the copies and a state of the art web-based interface that will permit users to view any number of quartos at once and create their own exhibits and annotations based on items they locate in the collection. In this paper I will discuss the challenges of using TEI P5 for a dynamic, interactive image-based edition.

We decided to use TEI for the Shakespeare Quartos Archive (henceforth called SQA) for the predictable reasons—the advancement of a the common standard with the concomitant potential for interoperability of the data with a variety of interfaces and the promise of data persistence across systems over time. We hoped that the page-image features of the P5 standard (e.g. facsimile, surface, and zones tags) would make up for what was sorely lacking in earlier iterations of the standard. We were, as anyone who has worked extensively with the standard might predict, disappointed but not despairing.

I do not regret the decision to use TEI P5 for the archive. There are, after all, few reasonable alternatives at present. However, it is clear that TEI has not yet sufficiently evolved the ability to encode data in images with even the flexibility of the early versions of HTML. In this paper I will discuss three functions in particular that proved unnecessarily difficult to implement: processing structurally coded XML for display in a page-centric interface, identifying regions within images, and integrating user-generated annotations into the data model of TEI.

My first complaint is more of a quibble than a real problem, but it is one I know many share. The current zone tag allows only for rectangular regions. Lou Bernard and others on the various TEI listservs have argued that this is all that is really necessary and that other XML schemas, like SVG, can be imported to handle anything else. I appreciate the desire for simplicity in the schema, but I question whether providing for a very limited shape (the rectangle) and then forcing users to import entirely new schemas to achieve the sort of functionality allowable even by HTML is really all that simple. I would disagree with, but could respect, a decision to leave shape descriptions out of TEI altogether; such a decision, at least, would be consistent with the stated goal of simplicity. But, surely, if the standard is going to natively support one and only one sort of shape, a set of three or more coordinate-pairs defining a polygon is a better choice. Of course, for curves and circles this might also be insufficient, but it is at least possible to approximate a circle with a series of points. I do not know how to make a circle with a square.

Additionally, as in so many projects, the problem of overlapping hierarchies has plagued the encoding and processing of the Shakespeare Quartos Archive. SQA presents the user with a desktop-like environment of draggable, resizable panels that can display, for a given page in a quarto, either a digital image or a textual transcript. The user, through the use of navigational buttons, can advance forward through the quarto, reading the page images or the transcripts in order. Such an interface is relatively common in image-based editions, but seems needlessly complicated to produce in TEI where pages, in order to solve problems of overlapping hierarchies, are usually represented as empty milestones. This is a familiar problem in TEI encoding, of course, and has proven so bothersome as to be the subject of a special interest group of the organization. Various attempts at solving this problem through segmented tags or concurrent hierarchies of various sorts are still being discussed, but the use of an empty milestone tag for page breaks is, while not a universally accepted solution, probably the most common one.

Milestones for page breaks make a great deal of sense when the page division is perceived by the audience as a vestigial interruption in the data stream, useful only for those interested in the physical properties of the analog format represented by the digital surrogate. However, in the case of an image-based edition, the page is in many ways the central unit of data organization. In order to return all of the relevant XML for a current page, and then transform this XML into HTML, it is necessary to suppress the text not between these breaks while preserving the structure of the XML document tree. This can be a resource intensive and time-consuming process which is not natively supported by XSL most DOM processors. Ultimately, we wrote a script to strip out all of the text nodes not located between the desired page breaks, find the lowest common ancestor of both page breaks, then transform this ancestor node into HTML using an XSL transformation. This process took far too long to be acceptable for the dynamic web environment we required, and so we finally resorted to preprocessing the XML into a series of HTML pages to be loaded as needed via AJAX calls.

Further, in the Shakespeare Quartos interface, we will allow users to add their own “tags” to regions of text or image. While this data will ultimately be stored in a database, it seems logical that the user-generated annotations should also be represented in TEI. Unfortunately, even the new P5 guidelines remain solidly rooted in the pre-web 2.0 world. The user/reader of a TEI document is allowed to passively absorb the text, but is not allowed to comment on it (at least in the same language as the text being described). Of course, some of this is built into the relatively static nature of XML on the web. XML files tend to be static; databases are what are modified. Still, without a standard way of representing annotations of TEI across systems, we are consigned to the same sorts of idiosyncratic, interface-specific solutions that TEI theoretically claims to overcome.

Of course, user annotations are not an easy thing to encode. My earlier problems with page breaks seem positively simple when compared with the conflicting hierarchies that necessarily ensue when we allow users to generate tags. Both problems force us to look hard at the solutions suggested by the members of our community concerned with overlap, and think very hard about the possibility of completely restructuring the language standards to account for our very non-hierarchical data sets. My personal preference is for standoff markup. In this method, the content is separated from the XML tree. The tags point to positions, labeled either through milestones in the content or through identifying the start and end offset of the range of characters that are described by a particular tag. The usual objection to this method is the very real problem that a single addition or deletion of a character in the text area will break the entire system.

To solve this, I recommend that the TEI community develop a TEI-compiler that turns “normal” XML into standoff markup. A TEI editor can write the mark up the text in the usual way, and then, as part of the validation/publication process, compile the XML. Any change to the document requires recompilation, of course, but I believe the slight slowing of the revision process may actually encourage archiving and preservation of earlier drafts. This, in turn, will make the citation of our online work an easier matter. Too often fellow scholars are forced to cite versions of our digital work that no longer exists because a quick update and “Save” obliterated the object of their reference. In the SQA, we will represent user annotations in structured XML that identifies the author of the comment along with whatever biographical information he or she chooses to provide, and then include a link to the file, offset, and length that is the subject of the note. This is only possible, because any published version of XML will persist even if the XML is later updated. If transcripts and XML were modified without cost or record, a user who annotates our texts could never be sure that her XML is pointing to the text she means to comment on.

The revisions to the TEI standards in the P5 update represent an important step in the right direction, but I believe they are no where near as expansive as they should have been to bring TEI into the 2.0 web that now serves as the primary distribution mechanism for our work. In the course of developing the SQA, we have been forced to cobble together non-standard methodologies to work with our TEI-compliant XML. It is my hope that we can design the next iteration of TEI with an eye towards true standardization.

Burying Dead Projects: Depositing the Globalization Compendium

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

Shawn Day
Digital Humanities Observatory
shawn@shawnday.com

Introduction

How do you end a project? Usually in our conferences we talk about imagining projects, starting them, getting funding, gaining respect, managing them and their research outcomes – but how often do we think about their ends, how we close them down and leave our digits behind for others. In this paper we propose to discuss the ends of a particular digital project, the SSHRC funded Globalization and Autonomy online Compendium. In the paper we will do the following:

1. Provide background to the project.
2. Demonstrate the major digital outcome of the project, the Globalization Compendium.
3. Discuss the problem of ending and how the project has been deposited.
4. And conclude by talking about the challenges of closing projects gracefully.

Project Background

The Globalization and Autonomy Online Compendium was one of two major coordinated outcomes of the Globalization and Autonomy project. The other outcome was a 10 volume academic book series being published by UBC Press. The project was supported by an MCRI grant of $2.5 million from the Social Sciences and Humanities Research Council that was awarded in 2002. [1]

The project was led by William Coleman at McMaster University, and involved over forty co-investigators in twelve universities across Canada and another twenty academic contributors around the world, not to mention the graduate students funded and staff. Unlike some projects, the digital component was woven in from the beginning. It was written in, budgeted, and reviewed. This included a review of the online Compendium by UBC press as part of the review of the book series. We also promised that the online component would be deposited and budgeted for that process. [2]

The core research objective of the project was “To investigate the relationship between globalization and the processes of securing and building autonomy.” The project was intended and administered to understand globalization in a collaborative and interdisciplinary way that avoided the often political and economic focus of globalization research. Hence the “autonomy” in the title - to explore globalization and resistance to globalization.

Demonstration

We will demonstrate the following aspects of the Online Compendium in order to illustrate some of the challenges faced when depositing it:

- Home page: we will discuss the scope of the Compendium

- Connection to Volumes: we will discuss how the Compendium and print volumes are connected

- Types of documents: we will show the different types of documents included and how they are linked. There are articles, position papers, summaries, glossary articles and a bibliographic database

- Articles: we will show how the design of the dynamically generated site allows different views to be generated from the raw XML to PDF

- Bibliographic database: we will discuss the issues around maintaining a central bibliographic database that is changed as papers are submitted and encoded

Digital humanists can probably imagine how we built this, but we need to review the structure of the project to understand what it is that we want to bury and the challenges inherent in the deposit process. Most of the content in the Compendium was written by participants and submitted as MS Word files for editing. We decided early not to force contributors to learn our XML scheme, a scheme which was developed through a consultation process based on the TEI guidelines. The editor and assistants carried out the encoding which also allowed us to add consistent links, especially for glossary articles of which there are over 200. We also used a custom search and replace batch tool to update the links regularly as we got more glossary entries.

The TEI XML files were parsed, verified, and processed when uploaded through the administrative interface. Metadata was extracted for the databases to generate tables of content dynamically and the files were indexed for searching. Bibliographic entries were the one form of content entered directly through a web interface. At an early stage, we identified the challenge of synchronizing a single bibliographic database for the project with all the bibliographies in the individual position papers, articles and glossary entries as they were submitted. As documents were submitted, the vagaries of individual authorial preferences, not unexpectedly, led to slightly different information for the same reference. We needed a way to normalize the bibliographies. What we constructed may seem Byzantine at first, but through application you will see how it makes sense.

1. When an editor encodes the bibliography of an article she doesn’t do it in the XML file. She enters any new records into the online database and checks any entries that already exist against the existing record.

2. If there is a discrepancy she follows it up and corrects the database entry if needed.

3. Once an entry is checked in the database she generates a stub tag with a key that corresponds to the database record. That is put into the XML file for the article rather than a full record.

4. When the XML version of the article is uploaded to the Compendium the system replaces the stubs with a full TEI `<bibl>` entry from the database. Thus each article has a full bibliography marked up in XML at the end should we deposit the articles independently of the bibliographic database. The Compendium project team felt this was important because they didn’t want their writings dependent on other data for completeness.

5. In the uploading we also keep online all the XML files with just the stubs. This allows us to periodically rerun the process that adds the full bibliography and replaces the full XML files thereby eliminating any inconsistencies that might occur as we correct entries over time. In effect we regenerate the content on a regular basis.

The Compendium itself in order to be experienced therefore is a system with the following components:

- XML files of the content

- A MySQL bibliographic database

- A metadata database of the content for generating pages and for searching

- A full text index for searching the text
- The code that handles the dynamic generation of the site, the searching, linking and the XSL transforms.
- Some HTML pages, and CSS stylesheets
- And various images that are embedded in pages.

Depositing What?

The XML files are not the site, and therein lies the problem of burying the whole Compendium. The experience of the Compendium is not only in the individual articles, or even in the bibliographic data – it is in the interaction between these mediated by the code and in the user experience. The glossary is a prime example – the meaning is not just in the text of entries, but in the searchable whole and web linking articles to others. At the start we thought it would be trivial to deposit the Compendium. We promised that we would encode the content following TEI guidelines and then deposit it at the Oxford Text Archives and other similar digital archives, but of course, the XML is not the Compendium. The Compendium is a work of its own that is more than the sum of the XML files. How do we deposit a system?

Why do we make such a big deal about ending the project and what exactly is the problem? The simple answer is that the funding body requires projects to deposit their research data. To quote from the SSHRC Research Data Archiving Policy, "All research data collected with the use of SSHRC funds must be preserved and made available for use by others within a reasonable period of time." If you accept the research funds you are obligated to do so, even though in a survey SSHRC conducted as part of the National Research Data Archive Consultation they discovered that very few had.

Another answer is that this is what we should do. Projects should be designed from the beginning to die gracefully leaving as a legacy the research data developed in a form usable in the future. That is what scholarly encoding following best practice guidelines like the TEI is about – encoding your data so that others can understand the decisions and be able to reuse it. We are fooling ourselves if we think projects will survive over time as living, well maintained projects. Ask yourself how many projects you have buried without a service.

SSHRC provides some of the reasons for archival deposit:

> Sharing data strengthens our collective capacity to meet academic standards of openness by providing opportunities to further analyze, replicate, verify and refine research findings. Such opportunities enhance progress within fields of research as well as support the expansion of inter-disciplinary research. In addition, greater availability of research data will contribute to improved training for graduate and undergraduate students, and, through the secondary analysis of existing data, make possible significant economies of scale. Finally, researchers whose work is publicly funded have a special obligation to openness and accountability. [3]

They don't say it explicitly, but one reason to deposit is that our research itself is of its time and grist for the mill of future researchers who may want to study us. Funders expects us to be open so that others can study the research process once we are dead, buried and history – a rather sobering prospect, but one of the features of an emerging philosophy of open research that advocates for exposing the research process rather than hiding the mess behind authoritative results.

Depositing the Compendium

The challenge then is determining what exactly are we to deposit and where? We conducted an environmental scan to attempt to identify best practices and ascertain what others have been doing to address this digital project archival challenge. Our research and recommendations were compiled and made available publicly on the TADA (Text Analysis Developers Association) wiki. [2] We should however warn that like all wikis this is a working document that was written during the process and has numerous dead ends or rough notes.

We identified the following separate types of knowledge that we might try to deposit:

- **Content** – this is the obvious one – the original research articles encoded in XML and other documents created by the project.
- **Code** – also obvious, but difficult. Why do it? Because in the code lies the interactivity and interface. This includes the XSLT code that generates the interface. The point, however, is that only with the code could one recreate the site as a dynamic site.
- **Process** – this is even less obvious, but the Compendium is the result of various research, programming and editorial processes, many of which are documented in instructions to authors and coders and other administrative documents, including documentation around the deposit process itself like the wiki mentioned. The process whereby we han-

dle synchronizing bibliographic entries is a case in point – it made a difference to the content and isn't apparent in the final XML files which hide the process. These materials document the Compendium as a collaborative writing project.

- The **Interactive Experience** – ultimately, for reasons that will be clear below, we also document and deposit information about the experience of the Compendium as an interactive work so that a future user could imagine what it was like to use the Compendium online. We chose to document this with a narrative with screen shots of typical use of the Compendium.

The Solution

So what was the recommended solution?

1. First, we decided to deposit these four components separately – content, code, process, and experience, each in the best format we can find. This is easy for the content, we designed it from the start in XML which is, to a certain extent, self-documenting, but in the case of code it is less clear.

2. All the materials are output to a flat-file format, by which I mean that things like the bibliographic database will be output to XML. The code is commented so that it could be compiled and documentation generated in HTML or XML in an industry standard fashion, although we must note, these standards are for documentation, not preservation. The point is that the documentation will be embedded in the code and could be extracted to produce documentation assuming that future computer scientists recognize how to extract documentation.

3. We are also creating "READ-ME" documents describing the environment and arrangement needed to run the code. We realize this means we are not depositing a working system that someone could download, install, and run to recreate the Compendium. The databases are not being stored in their native format, they will have to be regenerated and we are not creating a compressed file of the whole site. Instead we are trying to deposit something that could be used to reliably recreate the current Compendium. The reason for this is simple, over the long term the chances that someone can recreate the hardware and software platform on which an installation could work will approach zero. Therefore, we are better off providing something that can be understood and reprogrammed, if needed, than something that can't be install anyway. A further reason is, as noted above, that the purpose of depositing is not only so people can recreate the original site, but also so they can study the Compendium and reuse it in unanticipated ways.

4. I should add that we also not trying to deposit something in a form that the interactivity could be maintained. There are models for preserving interactive objects so that they can be easily run on emulators. The most obvious is to move all the interactivity into XSL or other XML standards for interactive processing like SMIL. The reason we are not going that route is that it is too expensive, probably won't capture all of the interactivity, and we have little confidence in any of the candidate standards. Does anyone remember HyTime? It was supposed to do for hypermedia what the TEI did for texts. Instead we are depositing a description with screen shots of the experience of using the Compendium so that someone who also had the code and content could at the very least understand the experience, and if they chose to reimplement our code they could recreate the experience.

5. Third, we will deposit the materials in multiple forms including a printout on archival grade paper, multiple copies of archival grade CDs or DVDs, and direct data deposit to online depositories.

6. Fourth, we will deposit the materials at multiple depositories from the McMaster University Library to the National Library of Canada which is in the process of implementing the National Data Archive Consultation reports.

We believe that by depositing in multiple forms, in multiple locations, and with rich documentation of the process and experience we will have buried the Compendium in suitable open casket, ready for reanimation or reuse. Let the worms loose. [4]

Conclusion

This brings us back to graceful ends. There are other ends to the Compendium than its deposit. One end, that is not in the scope of this paper, but may be of interest is managing the review of the Compendium so that those of us who worked on the digital design, but are not published in the book series are recognized. When the project negotiated with publishers for the print volumes, one reason we selected UBC Press was that they agreed to conduct a peer review of the Compendium along with the print series. We often talk about how digital work isn't reviewed, which causes trouble in the academy, but in this case the Compendium was reviewed and, in effect, accepted for publication as a companion to the print series. We therefore have an obligation to the Press to deposit the Compendium, an unanticipated side-effect.

The end of a project is not the death of the research. It is tempting to try to extend the life of a project beyond its initial funding as online materials can be so easily updated and added to. We suspect many projects don't deposit their work because there is always more to be done and the possibility of another grant. The result can be a long and drawn out death where nothing is saved because any day the project could be brought back to life. We have found it strangely liberating to run a process of deposit and believe it can actually release the research to imagine new projects. What is missing is stories about how you can reasonably deposit a project even if you intend to continue the research. We will conclude by briefly describing how the Compendium is the seed for a Global Globalization Research Dialogue project that is proposing to build a social network of researchers that includes researchers in countries affected by globalization, but excluded from research circles. Burying the Compendium allows us to imagine a new project with a life of its own.

Notes

[1] See www.globalautonomy.ca for the Globalization Compendium and details about the project.

[2] This deposit process is openly documented at *Archiving the Compendium*, http://tada.mcmaster.ca/view/Main/ProblemOverview

[3] See the *SSHRC Research Data Archiving Policy* at http://www.sshrc.ca/web/apply/policies/edata_e.asp

[4] It should be noted that at the time of writing this proposal we have not completed the process, but anticipate that the final deposit will take place assuming our partners like the National Library have suitable depositories in place in time. We note, however, that when you search for depositories, there are precious few in operation. The Oxford Text Archive, for example, seems to no longer take submissions.

T-REX: a Text Analysis Research Evaluation eXchange

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

Stéfan Sinclair
McMaster University
sgsinclair@gmail.com

J. Stephen Downie
University of Illinois
jdownie@uiuc.edu

Introduction

A common complaint about tool development in the digital humanities is that scholars appear to be reinventing the concordance over and over. The concern is that there is no coherent venue where tools, methods, and algorithms are compared, improved and documented. One model that has been discussed, and to some extent pursued, is the development of formalized mechanisms for the peer review of tools [1]. Competitions and exchanges are another way of advancing the field. In the spring of 2008, the Text Analysis Developers' Alliance (TADA) organized a digital humanities tools competition called T-REX [2] to assist the digital humanities community in its efforts to make meaningful technological and scholarly advancements. Based on our T-REX experience and that of MIREX, we will present the case that competitions and exchanges can provide an engaging alternative to peer review for methodological advancement. The paper will:

1. Discuss the first competition run in 2008 and highlight some of the recognized contributions that are being presented separately as a poster collection.

2. Discuss similar competitions and exchanges, especially the Music Information Retrieval Evaluation eXchange (MIREX) and how they work to build a community of developers who advance a field.

3. Present the evolution of the next T-REX 2 competition and describe the process whereby the community can build a shared agenda for development.

The 2008 TADA Research Evaluation eXchange (T-REX)

The first T-REX was designed to evolve into a model like MIREX (Music Information Retrieval Evaluation eXchange) [3] and TREC (text retrieval domain) [4], see below. The first T-REX was therefore designed to seed the community with starting ideas for competition, evaluation and exchange so that a community could form. The response to T-REX was positive and among the many submissions received, judges selected winners from the following categories:

- Best New Web-based Tool
- Best New Idea for a Web-based Tool
- Best New Idea for Improving a Current Web-Based Tool
- Best New Idea for Improving the Interface of the TAPoR Portal
- Best Experiment of Text Analysis Using High Performance Computing

The categories were deliberately chosen to cover not only working tools, but also ideas, designs and preliminary experiments. A primary objective of T-REX is to encourage the involvement and collaboration of programmers, designers, and users. We wanted to get ideas for tools and extensions to tools as much as get tools submissions in order to involve a broad base. This approach also recognized the importance of prototyping ideas over the deliver of finished production tools.

In total we had 11 submissions from individuals and teams. The proposals were read and judged by a panel of three judges and the adjudication panel chose to recognize seven of the submissions as contributing to the imagination of the field. A poster session is being organized in parallel with this paper to show the seven recognized projects.

T-REX had a number of initial sponsors including the TAPoR project [5], SHARCNET [6], IMIRSEL [7], Open Sky Solutions [8] and the *Digital Humanities Quarterly* [9]. These sponsors have different reasons for participating, but all were interested in the long-term evolution of the competition. For example, SHARCNET, a high performance computing consortium in Southern Ontario is supportive of the competition because it is a way for the consortium to reach out to the humanities and they hope to be able to support competition activities specifically on high performance computing applications in the humanities. Likewise we hope to find a way to work with DHQ to circulate recognized ideas and competition-exchange documentation.

About MIREX (Music Information Retrieval Evaluation eXchange)

T-REX is modeling itself on MIREX [10], which was first run in 2005. MIREX 2005 comprised 10 different "challenges" (or "tasks" in the MIREX nomenclature) and evaluated 86 individual algorithm submissions. MIREX 2008 represented the fourth iteration of the event and evaluated 168 individual submissions divided over 18 different tasks. The two keys to the apparent success of MIREX have been 1) its bottom-up involvement of the broad MIR research community; and, 2) its integration into the annual "life-cycle" of MIR research publication.

Each Winter, a new MIREX wiki is made available where interested researchers post task proposals. Subcommunities of interest coalesce around many of these proposals (some also "wither on the vine"). The successful sub-communities engage in very lively debate about the nature of the proposed tasks. It is in these debates about task definitions that concepts are clarified and much of the real progress in MIR research can be seen to be made manifest. By late Spring, the task definitions have matured to include the creation of the common datasets, the evaluation metrics to be used, and the input/output formats for engaging the datasets. In the Summer, those mature proposals that have a minimum of three different participants are declared to part of MIREX. Participants then submit their algorithms to the IMIRSEL team at UIUC which are then run over July and August. To reinforce the "exchange" notion that underpins MIREX, each submitted algorithm must be accompanied by an extended abstract that describes the algorithm. By early Fall, the results of the task runs are returned to the participants who must then update their abstracts and have them posted on the MIREX wiki [11]. This timing is not arbitrary as it is designed to coincide with the annual meeting of the International Conference on Music Information Retrieval (ISMIR, the premiere conference in MIR). MIREX has a special and very important relationship with ISMIR. It is now established practice that MIREX be given a dedicated half-day of the conference schedule. This half-day includes a MIREX plenary meeting that brings participants and general community members together. It also includes a poster session devoted exclusively to the presentation of MIREX-evaluated algorithms. Participation in the MIREX poster session is the "cost of entry" and is mandatory. It is the combination of community debate and clarification, evaluation of results, posting of algorithm abstracts, plenary discussions and the poster-based interactions that make MIREX so effective in driving the growth and success of the MIR research agenda.

MIREX is a good model for T-REX because it is based in a arts computing community similar to that which has a stake in text analysis. MIREX has a formula that recognizes the needs of participants while also providing a framework for review and outreach to the larger community.

How T-REX will evolve and continue

T-REX is evolving from a competition to a "evaluation exchange" along the lines of MIREX. Competitions do not allow the community to negotiate the challenges of interest and then work towards them. Competitions do not really encourage the comparing and contrasting of algorithms. Instead competitions risk rewarding and promoting a small number of teams with the resources to create significant tools. For this reason we propose that the next round of T-REX will be structured more as a CE with the following sequence of events:

1. **Developing the next round of challenges**. The participants from the previous round (2008) and new interested parties participate in a round of discussions aimed at developing a consensus about specific text analysis challenges for the next exchange.

2. **Developing the training and text materials**. The competition/exchange administrators have to develop training and test materials for the new challenges.

3. **Invitations to the challenge**. Invitations and contest materials have to be circulated.

4. **Submissions gathered and tested**. The submissions have to be gathered and tested against the original challenges.

5. **Documentation of results**. The results of the tests have to be documented in a way that advances our knowledge. And then it starts all over.

What is new in the next round is the first two steps of developing a community that chooses the challenges rather than the T-REX team choosing them. By inviting participation in the development of the challenge categories the activity becomes more of an exchange of ideas about what should be done and what can be done. The experience of MIREX has guided us on this.

A major issue with competitions and exchanges is their administration. On top of significant administrative work, they need non-trivial technical resources and expertise. The *Ad-hoc Authorship Attribution Competition* run by Patrik Juola in 2003-4 is good example of the amount of effort that needs to be expended [11]. Juola put ~500-700 person hours into developing training and test materials and running his competition. Likewise MIREX devotes several thousand person-hours to develop training and test materials each year. Further Juola and MIREX have to run the submitted tools against the test materials and document the results in a way that helps all involved. The tools submitted, despite the most stringent criteria, never run just "out of the box." Solutions to these issues are being developed which include automated web service resources and novel community-based work distribution models.

Conclusions

Why run competitions or exchanges? The short answer is threefold:

1. It provides tangible evidence of what has been and therefore can be accomplished (i.e., helps overcome the "reinventing the concordance" problem)

2. It creates a community of inquiry that can focus on advancing the field together

3. It formally recognizes work done to advance the field and documents it

In other words competition-exchanges can serve to recognize work that is hard to recognize through traditional peer review mechanisms, especially design work that is not delivered in a production tool. Tool development would seem to be one of those fields where peer review is unlikely to work, partly because of the significant cost of reviewing code. A competition-exchange reduces the cost because it focuses each round on the algorithms and code for a particular and defined focus and involves the community in setting the focus. A competition-exchanges works more like a juried art exhibit with an annual theme where review happens communally. In effect a competition-exchange becomes a community of peers that manage review over time.

This paper will conclude with a call for participation in T-REX 2.

Notes

[1] http://tada.mcmaster.ca/view/Main/PeerReviewCluster

[2] http://tada.mcmaster.ca/trex/

[3] http://www.music-ir.org/mirex/2009/index.php/Main_Page

[4] http://trec.nist.gov/

[5] http://portal.tapor.ca

[6] http://www.sharcnet.ca/

[7] http://www.music-ir.org/evaluation/

[8] http://openskysolutions.ca/

[9] http://digitalhumanities.org/dhq

[10] http://www.music-ir.org/mirexwiki

[11] http://www.music-ir.org/mirex/2008/index.php/MIREX2008_Results

[12] http://www.mathcs.duq.edu/~juola/authorship_contest.html

Ubiquitous Text Analysis

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

Stan Ruecker
University of Alberta
sruecker@ualberta.ca

Peter Organisciak
University of Alberta
organisc@ualberta.ca

Stéfan Sinclair
McMaster University
sgsinclair@gmail.com

Introduction

One of the problems facing e-text content publishers and text analysis tool developers is how to connect the appropriate tools with content. Early usability studies around the TAPoR portal [1] suggest that having users think first of the tool and then of the text is to forcibly reverse the normal order of thought. Users do not think of tools to which they bring texts, but instead like to look at texts and call operations on what they see. Accordingly, in this paper we will do three things:

1. We will present the usability case for privileging texts over tools and presenting tools on the side, so to speak.

2. We will review various interface models developed by the TAPoR project and others for embedding tools into content interfaces.

3. We will review the challenges of connecting tools reliably to content, especially connecting tools to large digital library collections. In this context we will discuss technical and open source solutions to the connection issues.

The Usability Case

Humanists are used to looking at documents; they are not used to treating documents as tokens for processing by tools. Interfaces for text analysis like that prototyped in the Eye-ConTact project [2] that present a visual programming environment where processes are connected into a “pipe and flow” diagram are too abstract for most humanists. The TAPoR (Text Analysis Portal for Research) workbench model is arguably less abstract, but users work by selecting from a list of favorite texts and a list of favorite tools that they run the texts through, a process that effectively hides the texts. Usability interviews conducted by Wendy Duff at the University of Toronto Faculty of Information to help improve the portal interface [1] led us to add an "Analyze This" view that presents the text in one frame with appropriate tools in a separate frame on the same screen. This solution, however, is only useful where a user has gone to the trouble to set up an account and define texts to study. We believe that another promising strategy is to embed tools into environments that already have texts, where there is a lot of content already published dynamically and a tool panel can be added to enhance reading.

Demonstration of Interface Models

TAPoR has been working with a number of projects to provide embedded tools. One source of inspiration and background research are the reading tools provided in the Open Journal System of the Public Knowledge project [3, 4]. At this point in the presentation we will demonstrate the following:

1. The Toolbar in the Globalization and Autonomy Compendium <http://www.globalautonomy.ca/>. This was our first experiment with an embedded tool bar. The code is a long span of JavaScript, CSS and HTML that is placed in the stream that generates all text pages from research summaries to position papers. The tool bar appears discretely at the bottom of the right hand navigation bar and is collapsible. This is documented so others can use it, but unfortunately the code tends to conflict with other CSS and JavaScript so it has only been used on a few projects.

2. Digital Humanities Quarterly. For each tool in the TAPoR portal and likewise for each tool from TAPoRware (<http://taporware.mcmaster.ca>) we provide example code to allow people to easily drop a tool panel or drop-down menu that can call a tool. This is the model that DHQ adopted and a dropdown appears at the top of each article that transfers the user to the appropriate TAPoRware page with the appropriate URL inserted. Unfortunately the code is still complex and difficult to implement.

3. FlashTAT. In order to avoid the problem of lots of conflicting code we have been developing a YouTube-inspired Flash application called FlashTAT (for Flash Text Analysis Tool) that can be embedded with one <object> tag and which, because the interface is handled by the Flash application, does not conflict with existing CSS and JavaScript. This tool also has the virtue that it can link directly to results, in this case a list of high frequency words, so the user can see those results and play with them rather than having to invoke a tool to see anything. We believe this is one of the more promising approaches to providing content providers with an easy way to embed tool interface, though we haven't tested it extensively.

4. Digital Texts 2.0. Another and more mature approach is to experiment with emerging social plugin architectures. We are convinced that in the long run, especially for student and faculty portals (not to mention scholarly publishing portals) we need to have social tools that users can choose from and include in their personal study space. Stéfan Sinclair and Johnny Rodgers have developed with a FaceBook plug-in called Digital Texts 2.0 which gives users a social bibliography in FaceBook accounts.

The Challenge of Connecting Tools to Content

The challenge of such embedded tool projects is magnified if the tools are placed in large content collections. Even in our smaller experiments we have had to think about reliability and scale. Some of the challenges we are currently addressing include:

Content producers will not embed tools if they are not reliable and if they won't scale. Typically, research tool projects are not funded to run a large-scale service. One solution is to give content producers a path from experimental use, where the tool runs off our tool server, up to giving them the code and helping them set up and run their own tool server so that they can guarantee reliability, or at least respond when the tools don't work. One disadvantage of handing off the code is that it makes updating the tools difficult; another is that we can't centrally gather usage statistics.

Embedded tools, especially opaque ones that use Flash, are difficult to customize to the design of the site they are embedded in. A programmer comfortable with CSS and HTML can adapt the look of tool bars like the one produced for the Globalization Compendium. We have provided some parameters to FlashTAT that allow its size and colour scheme to be customized using a special CSS file, but that undoes the advantage of a strategy where one <object> tag gets you a tool bar.

Social plug-in models are not mature. The FaceBook architecture is proprietary and FaceBook is not really a content portal. Should Google's OpenSocial be widely adopted by providers of portal frameworks then it is possible that social tool developers could develop to one Application Programming Interface (API) and be available in multiple portals and social applications.

Differentiating content and tools can be important for scholarly work, especially for quoting results and citing resources. Although we generally want to embed tools as seamlessly as possible into content, it is also important to make clear the distinction between the two as users might want to integrate them differently into their research. The tool itself, when embedded, potentially becomes part of the content and could confuse other tools.

The most difficult challenge ahead, however, is in overcoming the differences between the digital library culture that mounts and maintains online text collections and the culture of text analysis tool development that is more of a research craft. We need to find venues for discussing what content providers want and connecting them with research developers in the community. In conclusion, we will demonstrate an experimental essay, "Now Analyze That" [5] which presents a different embedded tool paradigm where tools are woven right into the prose of an essay, allowing users to recapitulate analysis that led to claims in the essay. Such a model connects not to content providers so much as to research authors [6] and the model presents deeper challenges to tool developers.

Notes

1. Cherry, J., & Duff, W. "Studying the usability of TAPoR, A Text Analysis Portal for Research." Faculty of Information Studies, University of Toronto, Research Day, March 10, 2006.

2. See Rockwell, Geoffrey and John Bradley. "Eye-ConTact: Towards a New Design for Text-Analysis Tools." *CHWP* A.4, publ. February 1998. <http://www.chass.utoronto.ca/epc/chwp/rockwell/>

3. Open Journal System, Public Knowledge Project. <http://pkp.sfu.ca/?q=ojs>

4. Siemens, Ray et al. "A Study of Professional Reading Tools for Computing Humanists." A report at <http://etcl-dev.uvic.ca/public/pkp_report/> that has been submitted to DHQ for publication.

5. Rockwell, Geoffrey and Stéfan Sinclair. "Now Analyze That". <tada.mcmaster.ca/Main/NowAnalyzeThat>

6. Smith, Jeff. "Penelope: A Practical Creative Tool for Integrating Authorship, Annotation, Analysis and the Management of Ideas." Paper presented at the Canadian Symposium on Text Analysis (CaSTA) Conference: New Directions in Text Analysis. A Joint Humanities Computing, Computer Science Conference at University of Saskatchewan, Saskatoon, October 16-18, 2008.

Links

TAPoRware: <taporware.mcmaster.ca>

TAPoR Portal: <portal.tapor.ca>

Digital Texts 2.0: <tada.mcmaster.ca/Main/DigitalTexts2>

FlashTAT: <tada.mcmaster.ca/Main/FlashTAT>

Globalization and Autonomy Compendium: <www.globalautonomy.ca/>

OpenSocial: <code.google.com/apis/opensocial/>

Ubiquity <wiki.mozilla.org/Labs/Ubiquity>

Gen Y Teaching Gen Y

Meghan Rosatelli
Virginia Commonwealth University
rosatellime@vcu.edu

In the August issue of Adbusters Magazine, Douglas Haddow claims that the new, twenty-first century counterculture is the unfortunate hipster. Haddow claims, in no uncertain terms, that Gen Y is so unoriginal, narcissistic, and media obsessed, that we fail to even create a suitable counterculture. We are, in effect, doomed to the digital world that spawned us and doomed to the insignificance of a "metrical mass" where our unconscious is infinitely plastic and subject to the environmental forces that surround us (Borch-Jacobsen, 1982). The *Adbusters* article struck a chord, not because my students brought me both digital and hard copies of the magazine during the first week of class, but because I realized that Haddow was not just talking about my students—he was talking about me. My course, "English 391: Reading Counterculture: 1950-present"[1] is based on the interdisciplinary MATX PhD program and has a diverse student body ranging from the School of the Arts to English. All of the students are juniors or seniors, making them between twenty and twenty-three years of age. I am twenty-six. We are Gen Y, and we are getting slammed—our meticulously mismatched outfits and all.

This research attempts to answer questions that arose, and continue to arise, as I teach media and culture literature courses. How do I teach my own generation media and culture? How do I utilize technology in the classroom that is meaningful? And, how do I keep their attention— and my attention for three hours each week, and up to six more with homework and reading assignments? The first question skews the subsequent questions dramatically. Appealing to Gen Y and bringing relevant technology into the arts and humanities is, fortunately, a well-researched problem. David Buckingham's *Media Education: Literacy, Learning, and Contemporary Culture* (2003) identifies and explains strategies for incorporating and teaching media in the classroom in order to promote new sites of learning—a democratic educational approach made popular in the seventies and continued today. Joe Lockard and Mark Pengrum's (Ed.) *Brave New Classrooms: Democratic Education and the Internet* (2007), delve into tactics concerning dynamic teaching on the web. The editors also incorporate the pros and cons of digital education—namely the lack of corporeal awareness that is often subscribed to web users.[2] Tara McPherson, another editor of an inspiring collection of articles, culls a wide variety of educational/generational topics, such as current notions of temporality and the increasingly blurry divide between citizenship and consumption in *Digital Youth, Innovation, and the Unexpected* (2007). These publications, and the countless others not listed here, show the growing interest and relevant scholarship in the digital humanities and Gen Y. They are, for the most part, positive and forward thinking books and articles that attempt to utilize the unique skill sets of their students while transforming everyday technology into teaching tools. Yet, none of them specifically address the issue of the first digital generation becoming university instructors and professors. The top-down, rush to know the latest-digital-tools-in-order-to-appeal-to-a-newtype-of-student is slowly being pushed aside as a new generation of instructors appear on the scene.

As many of the aforementioned scholars note, I realized that incorporating technology into the humanities classroom was a tricky task, as was sharing my generational status with my students. My two biggest challenges came with utilizing technology in a way that appealed to my students (knowing the technology that was being "used" on me as a student), and teaching in a way that recognized, instead of ignored, our Gen Y status.[3] In short, the students and I had to be in it together—all in—or the entire course would collapse. And, I had to establish this "all in" concept early. After airing our mutual hatred of discussion boards, the tedium of being forced to respond to posts, and a host of other unfortunately typical uses of technology in the classroom, the students began keeping personal blogs. They had to post at least once a week (Saturday at midnight) and, initially, they had to utilize just about everything the blog had to offer in order to build a comfort level with embedding video, creating links, and working in a hypertextual fashion.

A common misconception permeates many "top down" views of Gen Y: we are all technically savvy. I can say with certainty, this is not true. The simpler the technology, the better, the more aesthetically pleasing the technology, the better, the more personal the technology, the better. I can also say that nothing irritates students more than to feel as though they are the guinea pigs in a technology experiment where the instructor is conveniently separate—viewing from above. So, I blogged, I linked, I embedded, and I read each link, watched each video, and listened to each song on their twenty-four blogs. The technology extended our conference style class discussions throughout the week, but the technology was not enough—and it never is. We critically assessed our Gen Y status as a perspective and a coda to the vibrant counterculture movements that preceded us. Instead of being bogged down by an often pejorative and inaccurate label, we interrogated our actions, views, and aesthetic choices. The result was a class that used simple and accessible technology in moderation, utilized various forms of media in the classroom to spur discussion (video, sound, etc.) and accepted, instead of ignored, our unique similarities as the "plugged in" generation. (I think, secretly, we each sought to prove Douglas Haddow wrong.)

The purpose of this research, as noted above, is to investigate the benefits and restrictions that generation on generation teaching has on both classroom dynamic and the uses of technology in the classroom. I believe that a shared generational status, although a fleeting moment in time, can provide new and interesting perspectives on teaching pedagogy in the humanities.

Works Cited

Borch-Jacobsen, Mikkel (1982). *The Freudian Subject*. Stanford: Stanford University Press.

Buckingham, David (2003). *Media Education: Literacy, Learning, and Contemporary Culture*. Cambridge: Polity Press.

Lockhard, Joe and Mark Pengrum, Ed. (2007). *Brave New Classrooms: Democratic Education and the Internet*. New York: Peter Lang Publishing.

Hayles, N. Katherine (2005). *My Mother was a Computer.* Chicago: University of Chicago Press.

McPherson, Tara, Ed. (2007). *Digital Youth: Innovation and the Unexpected*. Cambridge: MIT Press.

White, Michele (2006). *The Body and the Screen: Theories of Internet Spectatorship*. Cambridge, MIT Press.

Notes

[1] This course has a companion course "Reading Twentieth Century Popular Culture" that I also teach.

[2] Michele White and N. Katherine Hayles discuss these themes in their books *The Body and the Screen* and *My Mother was a Computer*, respectively.

[3] Unlike technical course in digital design or generally straightforward courses in composition and rhetoric, counterculture required a certain street credibility (for lack of a better word).

Towards an Interpretation Support System For Reading Ancient Documents

H. Roued Olsen
University of Oxford
henriette.olsen@classics.ox.ac.uk

S. M. Tarte
University of Oxford
segolene.tarte@oerc.ox.ac.uk

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

J. M. Brady
University of Oxford
jmb@robots.ox.ac.uk

A.K. Bowman
University of Oxford
alan.bowman@bnc.ox.ac.uk

Constructing readings of damaged and abraded ancient documents is a difficult, complex, and time-consuming task, often involving reference to a variety of linguistic and archaeological data sets, and the integration of previous knowledge of similar documentary material. Due to the involved and lengthy reading process, it is often difficult to record and recall how the final interpretation of the document was reached, and which competing hypotheses were presented, adopted, or discarded in the process of reading. This paper discusses the development of an Interpretation Support System (ISS), which aims to provide a system, which can aid the day-to-day reading of ancient documents, and in future other damaged documents, by keeping track of how these are interpreted and read. Such a system will facilitate the process of transcribing texts by providing a framework in which experts can record, track, and trace their progress when interpreting documentary material. Furthermore, it will allow continuity between working sessions, and the complete documentation of the reading process that has hitherto been implicit in published editions.

Introduction

The process of reading ancient documents is traditionally undertaken by an expert such as an epigrapher, palaeographer or papyrologist. The expert uses accumulated knowledge combined with external resources to piece together an interpretation of each ancient document. Such interpretation is often prolonged, and it can be difficult for experts to maintain a record of the interpretations made whilst undertaking their reading (Youtie 1963). This is important when disputing interpretations and sharing hypotheses with other experts, or pausing the reading of an ancient text and hoping to continue the same thought process at a later time.

The Image, Text, Interpretation: e-Science, Technology and Documents project (also known as eSAD: e-Science and Ancient Documents, http://esad.classics.ox.ac.uk), aims to use computing technologies to aid experts in reading ancient documents. The project is developing an Interpretation Support System (ISS) that can support the day-to-day reading and interpretation of ancient documents. This involves advanced IT tools that can aid the interpretation of damaged texts such as the stylus tablets from Vindolanda (http://vindolanda.csad.ox.ac.uk) and image processing algorithms to analyse detailed digital images of the documents (Tarte et al. 2008).

Background

Although Classics as a subject has made much use of information technology (see Crane 2008 for an overview), the use of IT to aid in the actual reading process of ancient documents is in its infancy. Terras (2006) developed a prototype system which demonstrated that it was possible to propagate plausible and useful interpretations of ancient texts, in a realistic timeframe. This used linguistic and palaeographic datasets to provide the "knowledge base" which could inform a decision making system to aid experts in reading texts.

Decision Support Systems (DSS) have previously been developed in the Department of Engineering Science at the University of Oxford to aid multi-disciplinary teams working with cancer patients in making decisions about their treatment (Austin *et al*. 2008). This system is based on a set of rules and allows experts to analyse and interpret digital images while recording decisions made about diagnosis and treatment, and suggesting possible next action steps.

Building the Interpretation Support System

The research presented here, though inspired by the above-mentioned medical application, shifts the focus from a Decision Support System to an Interpretation Support System (ISS). In contrast with medical practitioners, experts transcribing ancient documents do not make decisions based on evidence but instead create interpretations of the texts based on their perception. The ISS relies upon the idea that an interpretation is made up of a network of minor perceptions (percepts) ranging from low level percepts such as "these three line fragments are an incised stroke" to higher level percepts such as "these five letters can make up the word *'legio'*".

We want to make this otherwise implicit network of percepts explicit in a human-readable format through a web browser based application. To build an explicit network of percepts leading to an interpretation, we define an elementary percept as a region of an image that contains what is perceived to be a grapheme. The image can then be divided into cells where each cell is expected to contain what is perceived as a character or a space. This division of the image constitutes a tessellation. A single document might be tessellated in various ways and each of the tessellations might yield either an interpretation or a dead-end, but in both cases, the explicit network of percepts will document this.

The making of the tessellation, which in itself is an interpretive process, marks the boundary between lower and higher level percepts. The lower level percepts are based on physical identification of the features of the document (through the application of image analysis methods to detect features such as strokes); the higher level percepts (words, groups of words) work more towards gradually adding meaning to the transcription in progress. Ultimately, an interpretation can then be represented as a network of substantiated percepts, which will be made explicit through an ontology. Here an ontology is defined as a model of the concepts found in a text such as the concept of a word that contains several characters.

The ontology aims to make the rationale behind the network of percepts visible and thus expose both: (a) some of the cognitive processes involved in damaged texts interpretation; and (b) a set of arguments supporting the tentative interpretation. The system will use the ontology as a framework to assist the expert through the different levels of percepts ultimately yielding a final transcription. The transcription is a part of the overall edition of which there may be several and it will be formatted in EpiDoc style XML (http://epidoc.sourceforge.net/) allowing further interaction with other documents.

Building the Knowledge Sets

Much of the knowledge base that serves as justification for the commitment to a given percept during the interpretation process will come from the experts. However, letter frequency, word-and character-lists from documents such as the Vindolanda ink tablets will provide an invaluable source of information which can be used to generate the statistical likelihood of patterns in language and writing which may appear on the texts. We have taken a new approach to the XML encoding of the Vindolanda ink tablets based on contextual encoding (Hippisley 2005). The Vindolanda ink tablets have been encoded with EpiDoc standard XML to a very detailed granularity. The contextual encoding which is then imposed on the documents consists on encoding words, person names, geographical place names, calendar references and abbreviations. For example any instance of the word *pulli* (='chickens') in a document will be encoded <w lemma="*pullus*" n="1">*pulli*</w>. This encoding provides us with the information that the word *pulli* has the lemma *pullus* under which we can index this instance of the word and that this is the first instance of this lemma in the document. This information has been used to generate word frequency lists and is extremely useful as a part of a knowledge base to build the ISS on. Further knowledge bases will be generated from the marked up dataset, to provide uncertainty and character frequency lists. Additionally, further work will be undertaken with the experts to generate lists of common percepts and interpretation making processes. By encoding these in XML, the knowledge sets for the system will be in place.

Conclusion

The construction of an Interpretation Support System for ancient texts, although ambitious, will provide a useful tool for those experts who work on developing interpretations of damaged documents by facilitating and recording the evolving interpretation process. Additionally, by making explicit the percepts which trigger such transcriptions of ancient documents, we will further our knowledge of the reasoning process undertaken by experts in propagating readings of ancient documents. Furthermore, the successful development of an image and language based Interpretation Support System will provide a set of tools which can be adopted and adapted by other domains which rely on detailed analysis and interpretation of image based material.

References

Austin, M., Kelly, M., Brady M. (2008), "The benefits of an ontological patient model in clinical decision-support". In Fox, D. and Gomes, C. P. (eds), AAAI, pages 1774–1775. AAAI Press.

Crane, G. (2008) "Classics and the Computer: An End of the History". In Schreibman, S., Siemens, R., and Unsworth, J. (eds). A Companion to Digital Humanities. Blackwell Companions to Literature and Culture. Blackwell. http://www.digitalhumanities org/companion/view?docId=blackwell/9781405103213/9781405103213.xml &chunk.id=ss1-2-4&toc.depth=1&toc.id=ss1-2-4&brand=default

Hippisley, D. (2005) "Encoding the Vindolanda tablets: an investigation in contextual encoding using XML and the EpiDoc standards." MA Dissertation submitted for the MA in Electronic Communication and Publishing, School of Library, Archive and Information Studies, UCL.

Tarte, S. M., Brady, J. M., Roued Olsen, H., Terras, M., Bowman, A. K. (2008), "Image Acquisition and Analysis to Enhance the Legibility of Ancient Texts". UK e-Science Programme All Hands Meeting 2008 (AHM2008), Edinburgh, September 2008.

Terras, M. (2006)."Image to Interpretation: Intelligent Systems to Aid Historians in the Reading of the Vindolanda Texts". Oxford Studies in Ancient Documents. Oxford University Press.

Youtie, H. C. (1963): "The Papyrologist: Artificer of Fact". GRBS 4 (1963), p. 19-32.

Design as a Hermeneutic Process: Thinking Through Making from Book History to Critical Design

Stan Ruecker
University of Alberta
sruecker@ualberta.ca

Alan Galey
University of Toronto
alan.galey@utoronto.ca

In what ways can the process of designing be used simultaneously for creating an artifact and as a process of critical interpretation? Can new forms of digital objects, such as interface components and visualization tools, contain arguments that advance knowledge about the world? This paper addresses those questions, first by exploring theoretical affinities shared by recent design and book history scholarship, and then by connecting those theories to the emerging practice of peer-reviewing digital objects in scholarly contexts. The paper concludes by suggesting that new forms of scholarly creation, especially those emerging from the digital humanities, need to be understood within the epistemic contexts that design and book history have concurrently been modelling in recent years.

One longstanding tradition of design is to understand it as an invisible handmaiden to content, where form follows function, and the typography in a book, for example, becomes transparent to the reader (Bringhurst 2005). Good design in this school of thought is design that goes unnoticed. An alternative tradition treats design as creative expression, where the hand of the designer is evident and we see a style that can be associated with the person responsible (Rand 1985). A related variation, sometimes referred to as "critical design," treats design as a rejection of the first tradition, resulting in typography that is intentionally difficult to read, and chairs that no one can sit in (Dunne 2005). All three of these approaches to design have their place, and we would argue that each of them can legitimately be understood as a form of interpretation. However, we would also propose that there is another distinct possibility, where one of the goals of the designer has been deliberately to carry out an interpretive act while in the course of producing an artifact. As Lev Manovich has publicly phrased it (2007) "a prototype is a theory." One of the functions of the artifact then becomes to communicate that interpretation, and to make it productively contestable.

Manovich's assertion has a close counterpart in Bernard Cerquiglini's claim for textual scholarship that "every edition is a theory" (1999, p. 79). The symmetry of these two statements extends to much of design and book history as cognate but often separate fields. Yet design has become a preoccupation in book history since the 1980s, especially following D.F. McKenzie's influential work on the sociology of texts (McKenzie 1999; McGann 1991), which emphasizes both the centrality of physical design and manufacture, and the importance of collaboration between multiple agents in the construction of meaning in books and other textual artifacts. This attention to the design of objects as "expressive form" (McKenzie 1999, p. 9) is poised to extend into the study of digital objects, including electronic literature and video games. Although McKenzie suggested a natural extension of bibliography's analytical and interpretive methods to texts in all media, including film, sound recording, and electronic text, the digital object presents challenges to hermeneutic assumptions carried forward from the print-based bibliography of the past century. Anthony Dunne describes the interdisciplinary challenge well when he asks "How can we discover analogue complexity in digital phenomena without abandoning the rich culture of the physical, or superimposing the known and comfortable onto the new and alien?" (2005, p. 17). In contrast to digital text production and software design, we have a fairly well-defined understanding of the traditional roles of non-authorial agents in print and manuscript book production, such as scribes, binders, typographers, compositors, correctors, and illustrators. "The sociology of texts" names an interpretive orientation which embraces these agents' contributions to the traditionally authorial process of meaning-making. In essence, then, book history has embraced design as a hermeneutic process, but has done so using a print-based vocabulary inherited from bibliography. The challenge now is to understand the kinds of agency that produce meaning in digital objects, and to appreciate the critical potential of digital objects in terms limited neither to print culture nor to the human-factors utilitarianism of industrial design (Dunne, 1999, p. 21-42).

By understanding how fields like book history take the design decisions embedded in physical artifacts as interpretive objects, we can begin to see digital humanists' creation of new digital artifacts as interpretive acts. We believe that the theoretical questions and convergences described above are strongly relevant to the emerging area of peer-review, evaluation, and authorship status of digital objects. Just as the boundary between digital documents and software applications has become less distinct due to web technologies, so has the boundary between traditional scholarly monographs and digital objects such as the "interactive media submissions" solicited by *Digital Humanities Quarterly*. By recognizing that digital objects – such as interfaces, games, tools, electronic literature, and text visualizations – may contain arguments subjectable to peer-review, digital humanities scholars are assuming a perspective very similar to that of book historians who study the sociology of texts. In this sense, the concept of design has developed beyond pure utilitarianism or creative expressiveness to take on a status equal to critical inquiry, albeit with a more complicated relation to materiality and authorship.

If we take seriously the suggestion that a digital object can embody an argument, then it should be possible to apply to digital objects some of the standard criteria for reviewing arguments. For Booth et al. (2008), the three key components of a good thesis topic is that it is contestable, defensible, and substantive. To be contestable, the thesis must be trying to convince people of a position that not everyone already believes. To be defensible, it must be possible, given the right kind of argument or evidence, that members of a reasonable audience could be convinced to change their minds. To be substantive, the argument must be worth the time and effort it takes for the writer to make it and the reader to engage with it.

For a prototype, we propose that contestability might reasonably consist of the inclusion somewhere in the interface of either an old affordance, previously seen in other interfaces, but now done in a new way, or else a new affordance – one not previously seen. Defensibility might equate to user studies of performance or preference. For old affordances handled in a new way, the studies could be comparative. For new affordances, comparison is not really possible, but there are strategies that can be adopted, such as looking at what we have elsewhere called "affordance strength" (Paredes-Olea et al. 2008, Ruecker 2006). Whether or not a prototype idea is substantive is somewhat harder to determine. This is, however, equally true for conventional scholarship.

The question of authorship is another factor to consider in the adoption of peer review of digital objects. Unlike research results in the sciences, arts research is still frequently published by a single author. However, in the case of digital objects, it is rare for a single person to be responsible for the entire process of conceptualization, design, development, and testing (Sinclair et al. 2003). At what point is a contribution significant enough to warrant the digital equivalent of authorship? Who should be first author – the person who had the original idea, or the person who did the bulk of the design, or the person who did the programming? These are questions which, if asked within a book-history context, would resonate with Roger Stoddard's often-quoted assertion that "authors do *not* write books. Books are not written at all. They are manufactured by scribes and others artisans, by mechanics and other engineers, and by printing presses and other machines" (1987, p. 4; emphasis in original). Peer review of digital objects thus involves digital humanities in a kind of sociology of texts with respect to the re-evaluation of authorship, while also encountering new aspects of digital design such as fragmentariness, modularity, and interoperability.

Ideas about design thus enter the digital humanities from a number of directions, each bringing certain disciplinary predispositions with them. The goal of this paper is a synthesis of design and book history perspectives on the ethos of "thinking through making," which informs much digital humanities research and pedagogy generally. The digital humanities must not lose sight of the design of artifacts as a critical act, one that may reflect insights into materials and advance an argument about that artifact's role in the world. Design thus provides a lynchpin for theoretical questions that unite different fields.

References

Booth, Wayne G., Gregory G. Colomb, Joseph M. Williams. (2008). *The Craft of Research*. 3rd Edition. Chicago: University of Chicago Press.

Bringhurst, R. (2005). *The Elements of Typographic Style*. 3rd ed. Vancouver: Hartley & Marks.

Cerquiglini, B. (1999). *In Praise of the Variant: A Critical History of Philology*. Trans Betsy Wing. Baltimore: Johns Hopkins University Press.

Dunne, A. (2005). *Hertzian Tales: Electronic Products, Aesthetic Experience, and Critical Design*. 2nd ed. Cambridge, MA: MIT Press.

Manovich, Lev. (2007). Q&A session at the Digital Humanities (DH) 2007 conference. University of Illinois at Urbana-Champaign.

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

McKenzie. D.F. (1999). *Bibliography and the Sociology of Texts*. Cambridge: Cambridge University Press.

Paredes-Olea, Mariana, Stan Ruecker, Carlos Fiorentino, and Fraser Forbes. (2008). "Using an Affordance

Strength Approach to Study the Possible Redeployment of Designs for Decision Support Visualization." Paper presented at the 9th Advances in Qualitative Methods Conference 2008. Banff, Canada. October 8-11, 2008.

Rand, Paul. (1985). *Paul Rand: A Designer's Art*. New Haven, CT: Yale University Press.

Ruecker, Stan. (2006). "Proposing an Affordance Strength Model to Study New Interface Tools." Paper presented at the Digital Humanities 2006 conference, at the Sorbonne, Paris, France. July 5-9, 2006.

Stoddard, R. (1987). Morphology and the Book from an American Perspective, *Printing History*, 9: 4.

Sinclair, Stéfan, John Bradley, Stephen Ramsay, Geoffrey Rockwell, Ray Siemens & Jean-Claude Guédon. (2003). "Peer Review of Humanities Computing Software." Panel at The Association for Computers and the Humanities / The Association for Literary and Linguistic Computing (ALLC/ACH): The 2003 Joint International Conference. Athens, Georgia. May 29-June 2, 2008. Draft articles online at http://tada.mcmaster.ca/Main/PeerReviewCluster

PCA, Delta, JGAAP and Polish Poetry of the 16th and the 17th Centuries: Who Wrote the Dirty Stuff?

Maciej Eder
Pedagogical University, Cracow, Poland
maciejeder@ijp-pan.krakow.pl

Jan Rybicki
Pedagogical University, Cracow, Poland
jkrybicki@gmail.com

Towards the end of the 16th century – at a time when theological disputes between Protestants and Catholics degenerated into open and, often, bloody conflicts – Polish literature saw a curious phenomenon termed "the age of manuscripts:" for fear of religious repression from either side, authors preferred not to publish their works in print; instead, they left them in handwritten form. The texts then circulated in a variety of copies. Sooner or later, confusion had to set in: different manuscripts preserved to this day ascribe individual poems to different authors or simply list the texts as anonymous.

Thus, in the present state of knowledge on "the age of manuscripts," we are dealing not only with a number of literary texts deprived of their authors, but also with a number of poets to whom not even a single poem can be reliably attributed. The history of literary studies on Polish Renaissance and Baroque poetry and the resulting scholarly editions show that the canon of almost every single poet has been changing considerably from scholar to scholar; some poems have been "reliably" attributed to several poets at a time. Present-day literature of the subject even contains authors brought to life by the scholars themselves, as is the case with the "Anonymous Protestant" and the numerous attempts at combining him with various poets of the late 16th century.

Quite by chance, the fate of the Anonymous Protestant attribution case is associated with another and much more crucial attribution problem. It so happens that the same manuscript (National Library in Warsaw, Ms. BOZ 1049) contains, several dozen pages further, a collection of thirty-one poems signed with the name of Mikołaj Sęp Szarzyński, one of Poland's most eminent poets, the author of dark metaphysical poems. Little is known of his life: he died young in 1581 and the rest is hypotheses. His entire attributed output amounts to little more than fifty poems; if thirty more could be reliably attributed to

Szarzyński, this would indeed constitute a priceless addition. The problem is that the thirty poems are all very vividly erotic, and, as such, do not fit the established if only hypothetical image of the existentialist sufferer.

In fact, Szarzyński's heritage abounds in unsolved issues. His only known collection of poems was printed 20 years after his death; the manuscript containing the problematic texts dates back to when he was still alive; but, since no confirmed Szarzyński manuscript is known, forensic handwriting evidence is unavailable. What is more, interspersed with the erotic poems are six non-erotic ones of undoubted Szarzyński authorship and in his well-known mannerist style; also, they appear in the printed posthumous volume.

Those who adhere to the view that Szarzyński indeed penned the erotica point out that the heady mixture of the sacred and the profane was characteristic of John Donne and the other English Metaphysicals, undoubtedly Szarzyński's mirror-images in both poetic tone and significance for their respective national literatures. Other researchers invented a spiritual breakthrough for Szarzyński: after a period of light-hearted erotica, he was supposed to decide to foreswear, in his own words, "the world's enticing vanities" and to radically change his poetic diction. Their adversaries maintain that writing in two different poetic languages. The erotic poems are visibly written in Renaissance mode, while Szarzyński's proven works (including the six found among the erotica in the manuscript) bear significant traces of a later mannerist paradigm. In this view, such a shift would be hardly probabile for an artist who died so young: at this rate, Szarzyński would have had to become a consummate master while still in his early teens.

The dispute on the uncertain erotic poems has not abated since the discovery of the manuscript in 1891 and has involved many of the most eminent scholars in the field; and yet the above-mentioned historical and literary arguments have not been enough to settle the matter. Attributions based on rudimentary statistics of parts of speech, enjambments and word order (Wyderka 2002) or chi-square tests of word frequency (Fleischer 1988) have given equally ambiguous results. Also, since the difficulty in this case naturally stems from the small size of the samples at our disposal (the erotic poems themselves only amount to some 2700 words), it was interesting to see if more advanced statistical methods could help solve this most eminent and fundamental problem in authorship attribution in Polish literature.

Although no other candidates had ever been proposed, our analysis included a number of authors active at the time, such as Mikołaj Rej, Jan Kochanowski, Łukasz Górnicki, Kasper Twardowski, Hieronim Morsztyn, Szymon Szymonowic, Szymon Zimorowic, in full awareness of the fact that these are but theoretical possibilities. With the same caveat, we included the anonymous author of *Tymatas*, a single poem of the same period, on the off-chance that he might be the mysterious author of the erotic poems.

The material for this study, then, consisted of samples by the above-mentioned major Polish writers of the broadly-understood turn of the 16th and the 17th centuries, at least two per author; the sample sizes were kept as close as possible to that of the suspect collection of erotic poems.

Both sets have been subjected to testing by three different tools: Multivariate Analysis (including Cluster Analysis, Factor Analysis and Multidimensional Scaling), Burrows's Delta (including Hoover's and Argamon's significant modifications such as DeltaOz, resulting in the powerful set of Delta spreadsheets, Hoover 2004, 2004a, 2007, 2007a, Argamon, 2008), and the recent black box software, JGAAP 3.3.1 (Juola et al., 2006, 2008), still in demo version. For the first two tools, various combinations of 'culling,' wordlist lengths and primary and secondary test groups have been used; in the third, which presents a variety of 'events' and statistical method combinations, those deemed the most reliable by the makers of JGAAP were used (Juola, pers. comm.).

In the first stage of the project, the three methods were evaluated for their reliability with this particular material and for the best possible parameters and versions of the procedures. Multivariate graphs at various levels of most frequent words (from 200 to 1000) usually grouped individual authors correctly, placing together not only two samples from single longer texts, but, just as successfully, very different writings by the same authors. Usually the one notable exception was Twardowski, whose data points seem to reflect the numerous turbulences and moral and emotional breakthroughs the poet underwent in his tempestuous life. Results for Delta (and especially DeltaOz, which proved of the highest reliability here) were even more satisfactory, identifying correct authors in the known authors' group with a reliability of 13 out of 15 (ca. 87%) at 250 most frequent words. JGAAP 3.3.1 fared even a little better at 21/24 (87,5%) correct attributions, achieved with two options: Kolmogorov-Smirnov Distance and Manhattan Distance.

Having established a reasonable reliability of the three methods used (and a good consistency between their results), they were then applied to attribute the collection of erotic poems in question, using the same parameters that had produced the best results in the test runs.

In multivariate graphs, both MDS and FA refused to place the data point for the erotic poems anywhere close to any of the Szarzyński samples. Various other candidates came closer – usually different ones at different parameters, Twardowski being the only one to do so with any consistency, but only when his one volume of erotica, *Lekcje kupidynowe* ("The Lessons of Cupid") was used in the primary sample; in this case, an unusual amount of culling (at 40%) was needed to overcome the lexical and/or generic bias. The same effect was observed in linkage distances in Cluster Analysis. In DeltaOz, Szarzyński was never ranked better than third as a candidate, Kochanowski, Twardowski and Zimorowic usually being presented as less unlikely. JGAAP behaved in much the same way: not one set of the reliable parameters identified Szarzyński as the most plausible author.

The conclusion that stems from this analysis is very serious. It seems that while no candidate has been found, with any consistency, as the most probable culprit, the almost-accepted solution to what is the most significant attribution riddle in Polish literature – that Szarzyński is the one who wrote the collection of erotic poems – is put to reasonable doubt by the very consistent results of stylometric investigation presented here. This is further compounded by the fact that while three or four writers are shown to be more probable than Szarzyński by our analysis, they are much less so basing on historical and biographical data (including fellow erotica-writer Twardowski). It seems that – unless new texts are found – the room of Polish 16th/17th-century poetry must remain as untidy as it has been; that not all poems found there can be neatly pigeon-holed to the few authors of the era we now know by name.

References

Argamon, S. (2008). 'Interpreting Burrows's Delta: Geometric and Probabilistic Foundations,' *Literary and Linguistic Computing* 23(2): 131-147.

Burrows, J. F. (2002a). '"Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship,' *Literary and Linguistic Computing* 17: 267-287.

Fleischer, M. (1988). *Frequenzlisten zur Lyrik von Mikolaj Sep Szarzynski, Jan Jurkowski und Szymon Szymonowic und das Problem der statistischen Autorschaftsananyse*, Munchen: Sagner.

Hoover, D. L. (2004) **'**Testing Burrows's Delta,' *Literary and Linguistic Computing* 19: 453-475.

Hoover, D. L. (2004a) 'Delta Prime?' *Literary and Linguistic Computing* 19: 477-495.

Wyderka B. (2002). *Przedziwny wszędzie. O stylu Mikołaja Sępa Szarzyńskiego na tle tendencji stylistycznych poezji polskiego renesansu*, Opole: Wydawnictwo Uniwersytetu Opolskiego.

Juola, P., Noecker, J., Ryan, M., and Zhao, M. (2008). **'**JGAAP3.0 -- Authorship Attribution for the Rest of Us,' poster, Oulu: Digital Humanities 2008.

Juola, P., Sofko, J. and Brennan, P. (2006). 'A Prototype for Authorship Attribution Studies,' *Literary and Linguistic Computing* 21:169-178.

Translation and Delta Revisited: When We Read Translations, Is It the Author or the Translator that We Really Read?

Jan Rybicki
Pedagogical University, Cracow, Poland
jkrybicki@gmail.com

One of the first success stories of Burrows's Delta, that of the 'Englishing of Juvenal,' strongly points out that some translators, such as Johnson, translate other people's writing as if they were writing their own work, while others, such as Dryden, are 'able to conceal their hand' (Burrows 2002, 2002a). Previous stylometric studies of patterns in multivariate diagrams of correlation matrices derived from relative frequencies of most frequent words in character idiolects (Burrows 1987) in a number of originals and translations (Rybicki 2006, 2006a, 2007, 2008) have shown that while some stylometric tools can show the difference between translator and translator, the relationship between such patterns between original and translation is at best unclear – as unclear, one could say, as is the relationship between the lists of most-frequent-words in two (or more) languages, where one-on-one correspondences are rare if not non-existent.

In the first-mentioned pioneering study, Burrows compared translations by a variety of authors to their own writing – a very natural material for translations of poetry, itself very much the domain of poets. But translations of prose are usually done by men and women who often are not novelists in their own right; while there is a reasonable guarantee that a Dryden or a Miłosz translation, however 'unfaithful' or 'free', can still be good poetry, can one really trust that a novel translated into another language by someone who has never invented a plot line of his/her own can be at least remotely connected to the original in terms of such trifling features as… literary style? Or, in other words, whose style is it that we see in the translation: some foreign version of the original author's, or simply that of the much less talented translator? For, obviously, Kurt Vonnegut's requirement for a perfect translator ('that he or she be a more gifted writer than I am, and in at least two languages, one of them mine,' Vonnegut, 1991) is rarely met, and probably never in translations of prose. In fact, traditional (i.e. non-computer- and/or non-statistically-assisted) translation studies speak of 'inherent intersubjective processes,' which result in an 'unpredictability of the target style' (Wilss, 1996).

The above also signifies that any study venturing into these uncharted waters is deprived of exactly the ideal comparative material Burrows could use in his 'Juvenal;' and that the best one can count on are other translations by the same translator, or other translations of the same author by other translators; if in luck, one can sometimes use different translations of the same novel. In this context, the main question of the title boils down to whether different works of an author translated by the same translator are going to be stylometrically more similar to each other than to translations by that translator of other authors; and whether translations of one author by many translators are going to resemble each other rather than translations of other individual authors by these translators.

Two discrete material sets have been used in this study. The first is a collection of 21 Polish translations of English-language fiction by Rybicki, made between 1991 and the present, with two authors appearing more than once: John le Carré (5 different novels) and Douglas Coupland (3). The second consists of 10 Polish translations of Jane Austen's novels by 4 different translators, produced between 1956 and 2006, and compared to 5 other translations by the same translators (at least 1 by each).

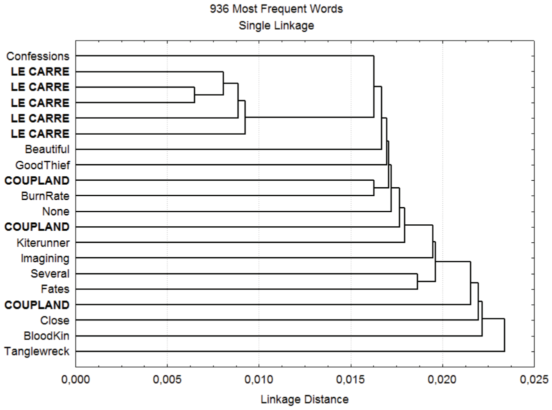


Fig. 1 Cluster Analysis of all Rybicki translations

Both sets have been subjected to testing by three different tools: Principal Component Analysis, Burrows's Delta (including Hoover's and Argamon's significant modifications such as DeltaOz, resulting in the powerful set of Delta spreadsheets, Hoover 2004, 2004a, 2007, 2007a, Argamon, 2008), and the recent black box software, JGAAP 3.3 (Juola et al., 2006, 2008), still in demo version. For the first two tools, various combinations of 'culling,' wordlist lengths and primary and secondary test groups have been used; in the third, which presents a variety of ‘events’ and statistical method combinations, those deemed the most reliable by the makers of JGAAP were used.

In the first set, all three methods clearly highlighted a great similarity of the five Rybicki translations of le Carré. Multivariate Analysis graphs (Cluster Analysis, Factor Analysis and Multidimensional Scaling) always contained a visible (yet not exclusive) cluster of this author (as exemplified in Fig. 1). A huge majority of various combinations of wordlist size, pronoun deletion, culling percentage, primary and secondary text selection, and Delta variety, correctly identified translations of le Carré from among all other Rybicki translations (the infrequent yet rule-proving exception being that of Barris, whose *Confessions of a Dangerous Mind* can be seen as a parody of the spy thriller genre, and who twice preceded le Carré, himself ranked second, as the most likely author of the original). In contrast, the three Coupland books exhibited much less stylistic consistency – a possible consequence of the significant time that had elapsed between the three translations. On the other hand, given the choice of only le Carré and Coupland as primary authors, Delta invariably identified all of their translated novels correctly and refused to mistake them for any of the secondary authors – and that at a very wide range of most-frequent-word lists. A very similar pattern emerged from some of the more robust statistical procedures available in JGAAP (word analyses with Kolmogorov-Smirnov Distance and Manhattan Distance).

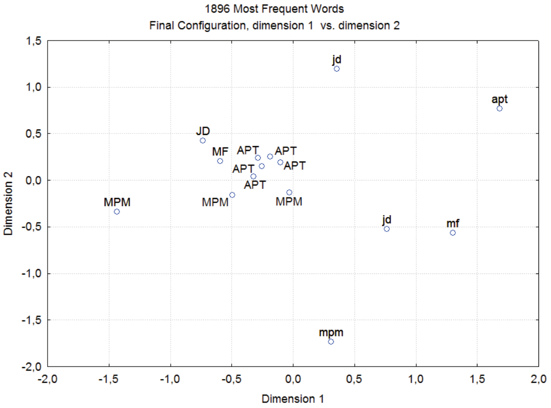


Fig. 2 Multidimensional Scaling Analysis of translations of Austen (uppercase initials) and of other original writers by the same translators (lowercase)

In the second set, multivariate graphs contained a visible cluster of all translations of Austen novels by all four translators, while their other translations hovered away from the cluster in other parts of the graph (Fig. 2). This was even more obvious in Delta analyses, where all Austen novels in Polish, irrespective of their translators, were almost always correctly identified as Austen novels (in 9 out of 9 cases at various combinations of parameters, Table 1). Conversely, standard Delta and DeltaOz never exceeded a 5/11 correctness ratio when asked to tell translator from translator (Table 2). Faced with an Austen text by translator A, JGAAP almost invariably chose a translation of any Austen novel by translator A or B, rather than a translation of another author by translator A, as the more similar of the two.

Words:	Correct	Incorrect	Ranked 2	Ranked 3	Ranked 4	Ranked 5+
100	9	0	0	0	0	0
200	9	0	0	0	0	0
300	9	0	0	0	0	0
400	6	3	3	0	0	0
500	5	4	4	0	0	0
600	9	0	0	0	0	0
700	9	0	0	0	0	0
800	9	0	0	0	0	0
900	9	0	0	0	0	0
1100	9	0	0	0	0	0
1300	9	0	0	0	0	0
1500	9	0	0	0	0	0
1700	9	0	0	0	0	0

Table 1. Correctness of Delta attributions of Austen novels in 9 Polish translations

	Correct	Incorrect	Ranked 2	Ranked 3	Ranked 4	Ranked 5+
100	1	10	3	4	3	0
200	3	8	3	4	1	0
300	3	8	2	5	1	0
400	4	7	2	4	1	0
500	5	6	0	5	1	0
600	5	6	0	5	1	0
700	5	6	0	2	4	0
800	5	6	0	2	4	0
900	5	6	0	4	2	0
1100	5	6	0	3	3	0
1300	5	6	0	1	5	0
1500	5	6	0	1	5	0
1700	5	6	0	1	5	0

Table 2. Correctness of Delta attributions of 11 translations to translators

The conclusions from this study are twofold. First, the results – reliable, because they were obtained using three tools using a variety of statistical calculations (traditional multivariate analysis, state-of-the-art Delta, and the emerging JGAAP black box with its various distances) – are an interesting addition to Burrows’s study of the Juvenal translations. The fact that translations of prose tend to ‘disguise’ the individual styles of the translators and seem to create individual styles of the original authors in a *foreign* language – styles common to a number of different translators even more consistently than translations of poetry – can be a consequence of the greater formal complexities of poetic translation, which usually calls for the task to be done by literary creators in their own right – exactly such as the translators studied by Burrows. Secondly, and perhaps more importantly, the fact that ‘unpredictable target style’ is in fact predictably similar between translations of the same original author done by many people – as measured on the basis of frequent and very frequent words (additionally culled to eliminate individual word-types characteristic for the vocabulary of individual novels and translators) – suggests possible and previously unsuspected affinities between original and translation at very 'mechanical' levels of the text. Indeed, since many of these words – especially those at higher frequencies – carry grammatical rather than lexical meaning, there is hope on the horizon that one day computational stylistics and cognitive linguistics (and particularly cognitive translation studies) might be two sides, experimental and theoretical, respectively, of the same phenomenon – as has already been suggested by Connors (2006, 2008).

References

Argamon, S. (2008). 'Interpreting Burrows's Delta: Geometric and Probabilistic Foundations,' *Literary and Linguistic Computing* 23(2): 131-147.

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford: Clarendon Press.

Burrows, J. F. (2002). 'The Englishing of Juvenal: Computational Stylistics and Translated Texts,' *Style* 36: 677-99.

Burrows, J. F. (2002a). '"Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship,' *Literary and Linguistic Computing* 17: 267-287.

Connors, L. (2006). 'An Unregulated Woman: A Computational Stylistic Analysis of Elizabeth Cary's *The Tragedy of Mariam, The Faire Queene of Jewry*,' *Literary and Linguistic Computing* 21(Supplementary Issue): 55-66.

Connors, L. (2006). 'Function Word Analysis and Questions of Interpretation in Early Modern Tragedy,' Oulu: Digital Humanities 2008.

Hoover, D. L. (2004) **'**Testing Burrows's Delta,' *Literary and Linguistic Computing* 19: 453-475.

Hoover, D. L. (2004a) 'Delta Prime?' *Literary and Linguistic Computing* 19: 477-495

Hoover, D. L. (2007). 'Corpus Stylistics, Stylometry, and the Styles of Henry James,' *Style* 41(2): 174-203.

Hoover, D. L. (2007a). 'Quantitative Analysis and Literary Studies,' *A Companion to Digital Literary Studies*, Oxford: Blackwell, 517-33.

Rybicki, J. (2000). *A Computer-Assisted Comparative Analysis of Henryk Sienkiewicz's Trilogy and its Two English Translations*. PhD thesis, Kraków: Akademia Pedagogiczna.

Rybicki, J. (2006). 'Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations,' *Literary and Linguistic Computing* 21(1): 91-103.

Rybicki, J. (2006a). 'Can I Write like John le Carré?' Paris: Digital Humanities 2006.

Rybicki, J. (2007). 'Twelve Hamlets: A Stylometric Analysis of Major Characters' Idiolects in Three English Versions and Nine Translations,' Urbana-Champaign: Digital Humanities 2007.

Rybicki, J. (2008). 'Does Size Matter? A Re-evaluation of a Time-proven Method,' Oulu: Digital Humanities 2008.

Vonnegut, K. (1991). *Fates Worse than Death*, New York: Berkley Trade, 181.

Wilss, W. (1996). *Knowledge and Skills in Translator Behaviour*, Amsterdam: John Benjamins, 34-5.

Juola, P., Noecker, J., Ryan, M., and Zhao, M. (2008). **'**JGAAP3.0 -- Authorship Attribution for the Rest of Us,' poster, Oulu: Digital Humanities 2008.

Juola, P., Sofko, J. and Brennan, P. (2006). 'A Prototype for Authorship Attribution Studies,' *Literary and Linguistic Computing* 21:169-178.

Collective Culture and Visualization of Spatiotemporal Information

Shinya Saito
Ritsumeikan University
saitos@fc.ritsumei.ac.jp

Shin Ohno
Ritsumeikan University
ohnoshin@gmail.com

Mitsuyuki Inaba
Ritsumeikan University
inabam@sps.ritsumei.ac.jp

Introduction

Archiving with digital technology is indispensable for historical cultural heritage items, but we also have to apply this technology for culture that remains difficult to form and store. This paper will argue three important aspects of archiving regional culture, its representation, sharing, and analysis. Then, in order to deal with these aspects, we will introduce the environment KACHINA-CUBE (KC) which we are developing and evaluate findings of its application. We will start with elaborating its design concept, then architecture and application, referring to its related works.

Design Concept

In order to design KC as a tool for archiving regional culture, we see the following three aspects important, thus take them into consideration.

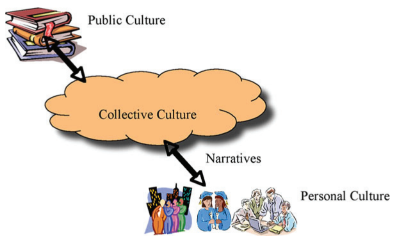


Fig. 1 Three levels of cultural representation

Representation Aspect

Developing Valsiner's concept of culture (Valsiner, 2007), we argue that cultural representations appear in three levels; personal, collective, and public (See Fig. 1). The first, personal culture consists of personal memory and knowledge. The second level of collective culture can be considered collection of personal cultures. The third level of public culture is public information, publicly recognized and found in textbooks and dictionaries. Public culture is well preserved. Because of complexity and ambiguity of personal and collective cultures, however, researchers are still struggling with how to preserve these two, which should be understood by putting them in their socio-cultural contexts. Taking a socio-cultural approach, Wertsch (1998) advocates to treat narratives as artifacts, the key to represent culture.

Sharing Aspect

W3C puts tremendous efforts to create standardized frameworks for Web, and researchers in digital humanities regard semantic web technology as one of the key research fields. This kind of technology gives us various chances to share data for other use. We believe archived cultural data should be standardized to fit in this framework, which allows users to access data and utilize them in various platforms.

Analytical Aspect

To research history and culture in a specific region, oral history plays an important role to make us understand them. Collection of oral history in a specific area reveals what kind of life experience people had and/or have in the area, their similarities and differences.

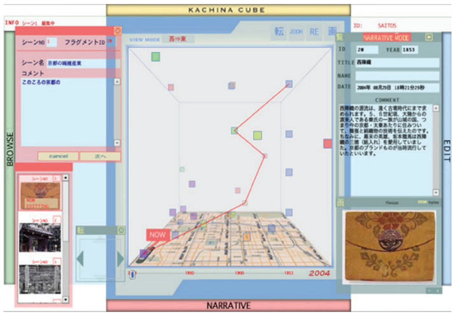
Valsinar and Sato (2006) propose the concept of Trajectory Equifinality Model (TEM) as a framework to analyze personal experience which suggests diverse and possible trajectories based on three concepts: Bifurcation Point (BFP), Obligatory Passage Point (OPP), and Equifinality Point (EFP). BFP is a point of each person's behavior branching or forking into new types of behavior by his or her choice. OPP means, literally, Obligatory Passage Point which most of the people have to go through because of their own logic, institutions, and customs. EFP is defined as the final state that individuals equally reach from different initial conditions. Focusing on OPP and EFP to investigates cultural recognition that people in a particular region share, this paper describes the KC system that assists users to seek through multiple narratives and identify OPP/EFP in construction of the regional culture.

Architecture

Giving consideration into these three aspects, we have been developing KC system. As for the representation aspect, we accept Wertsch's propose and design the software to hold data of spatiotemporal information.

We decided to design KC in three dimensions, two dimensions for geographical information and another one for temporal information. In this virtual 3D space (CUBE model) (see Fig. 2), users can post formal and informal story fragments. Among them, we call formal ones history fragments, and informal fragments story ones. KC also supports researchers to make linkages among fragments in periodical or logical order. We call a set of cultural fragments storyline.

As for the sharing aspect, we apply RDF/OWL to define our data. Its extensive and flexible definition is suitable for our system and motivates other researchers to access our data (Bray, 2001). We defined the data format as follows:



INFO シーン1 編集中

シーンNO 1 フラグメントID 20

シーン名 京都の繊維産業

コメント このころの京都の

cancel 次へ

ID: SAITOS

NARRATIVE MODE

ID 20 YEAR 1853

TITLE 西陣織

NAME

DATE 2004年 08月29日 18時21分29秒

COMMENT

西陣織の源流は、遠く古墳時代にまで求められます。5、6世紀頃、大陸からの渡来人である秦氏の一族が山城の国、つまり今の京都・太秦あたりに住みついて、養蚕と絹織物の技術を伝えたのです。ちなみに、幕末の英雄、坂本龍馬は西陣織の三徳（紙入れ）を愛用していました。京都のブランドものが当時流行していたといいます。

Fig. 2 Image of CUBE model

1. History fragment class: Objective information in textbook or dictionary
2. Story fragment class: Subjective information such as oral history
3. Storyline class: Aggregate of historical and story fragments based on a specific context
4. Geography class: Geographical information of the historical and story fragments
5. Temporal class: Time when the incidents told in historical and story fragments occurred

Finally, the analytical aspect, KC implements the function of OPP/EFP detector. OPP/EFP detector searches fragments occurred in similar places or time. Using OPP/EFP detector makes it possible to learn spatiotemporal possibilities. This is the implementation of Sato's concept, and to understand a region in meta level, this feature can be a strong analytical tool.

Application

As a test case, we applied the data of movie culture in Kyoto Rakusai Area, a.k.a. Japan's Hollywood, to our system. We used oral history data collected by Tomita and Itakura (2001). Each story that has spatial and/or temporal information is stored to the system. Currently we have oral history data from three storytellers who had involved in movie industries in the area from 1910's to 1930's. In terms of the representation aspect, the story fragments were well mapped on the 3D space, with their spatiotemporal information. Using our storyline representation visualizes connections among independent fragments. As for the sharing aspect, we are still working on definition of the data, hoping that this feature will be available soon. As for the analytical aspect, OPP/EFP detector displays different storytellers' worlds with possible alternatives experiences (see Fig. 3).

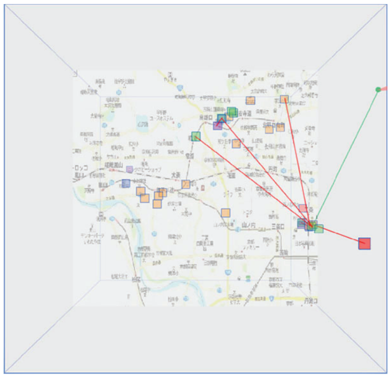


Fig. 3 Display of OPP/EFP detector

Conclusions

In this paper, we argue importance of not only archiving collective culture but also standardized semantic Web as a socio-cultural analytical tool which allows researchers to access data and utilize them in various platforms. Based on this argument, we developed KC, applying it to actual research. As a result, our system demonstrates a lot of potentials for research in various fields, which we have to prove by developing further this software with applications, as well as examining it in more case studies of collective culture.

There are well-known research projects and Web systems to deal with spatiotemporal information. For example, the TimeMap Project in the University of Sydney develops Web GIS that can visualize chronological data and animate historical maps (Johnson, 2004). The SIMILE Project in the Massachusetts Institute of Technology develops the TimeLine system that can organize text and pictorial data in chronological order. Our KACHINA CUBE is significantly different from these previous Web systems in the following two points:

1. Adoption of CUBE model (3D viewer that combined map with timeline); and

2. Implementation of user interface suitable to contain narratives and oral histories

With KC, therefore, we hope to contribute to further development of regional archive system and digital humanities in general.

References

Bray, T. (2001). *What is RDF?*. http://www.xml.com/pub/a/2001/01/24/rdf.html (accessed on 14 November, 2008).

Tomita, M. and Itakura, F. (2001). Voices from Kyoto: Interview with ITO Asako – An aspect of Japanese film history. *Art Research*. 1(1): 127-138.

Valsiner, J. (2007). PERSONAL CULTURE AND CONDUCT OF VALUE, *Journal of Social, Evolutionary, and Cultural Psychology*. 1(2): 59-65.

Valsiner, J. and Sato, T. (2006).Historically Structured Sampling (HSS): How can psychology's methodology become tuned in to the reality of the historical nature of cultural psychology? In Straub, J. Et.al. (eds.) *Pursuit of meaning. Advances in cultural and cross-cultural psychology.* Bielefeld: Transcript Verlag, pp.215-251.

Wertsch, J. V. (1998). *Mind As Action*. NewYork: Oxford University Press.

Johnson, I. (2004). Putting Time on the Map: Using TimeMap for Map Animation and Web Delivery, GeoInformatics.

SIMILE TimeLine. http://simile.mit.edu/timeline/ (accessed on 14 March, 2009).

The Apex of Hipster XML GeekDOM: TEI-Encoded Dylan"1: Understanding and Reaching a Community of Practice (A Case Study)

Analysis:
Lynne Siemens
Faculty of Business/School of Public Administration
University of Victoria

Ray Siemens
Faculty of Humanities
University of Victoria

Widget and Other:
Hefeng (Eddie) Wen
Dot Porter
Liam Sherriff
Cara Leitch
Karin Armstrong

Introduction

As the Digital Humanities/Humanities Computing community continues to develop more digital resources and accompanying tools, it is facing new challenges in regards to creating awareness and gaining acceptance of these among potential users. In particular, research projects must develop ways to disseminate information about the resources and tools to ensure that they are used and, at the same time, work to understand the community of practice which uses their resources. Viral techniques and other internet analytical tools present one option to achieve those objectives.

Context

As evidenced by the growing number of conferences, publications, and digital resources and tools themselves, Digital Humanities/Humanities Computing has been successful in their efforts to create useful resources and tools. As the field moves forward, in addition to responding to new technological challenges, community members must address new ones, often beyond their typical skills, capabilities and methodologies. One issue faced by many projects is the need to actively create awareness of these digital resources and understand their community of practice.

Among their recommendations, the LAIRAH project team outlined that the ideal digital humanities project has a clear understanding of their users and their needs as well as actively disseminates information about itself within the discipline and the community as a whole (Warwick, Terras, Hunginton, Pappa, & Galina, 2006) This is further reinforced by a recent report focusing on sustainability and revenue models for online resources. The authors discuss the importance of knowing one's users and their behaviours in regards to online resources. They go on further to highlight the importance of specific activities to create awareness and usage of these resources (Guthrie, Griffiths, & Maron, 2008). At a more practical level, Cohen and Rosenzweig (2005) devote an entire chapter in their book on digital history to the topic of building an audience for a digital resource. Various associations within this community are also exploring ways to increase the awareness of their activities while developing their membership base.

One tool often used by business to promote their products and services is viral marketing or electronic word of mouth. Unlike traditional techniques, viral marketing campaigns draw upon people's social networks and relies on individuals to forward on messages to others within their email lists (Dobele, Toleman, & Beverland, 2005; Smith, Coyle, Lightfoot, & Scott, 2007). For many organizations, drawing upon these individuals can also help build the products and services provide by the organizations, as Mozilla and Intuit have found (Cook, 2008; Freedman, 2007) . While this form of marketing is often used by companies to extend their reach with their customers and find new ones, it can be a useful tool for other types of organizations to define their present and potential communities of practice. But how best can the Digital Humanities community use this tool to its advantage?

Methodology/Steps

To determine the effectiveness of this tool and accompanying internet analytical tools within this community, a viral experiment was designed to showcase TEI and novel ways that it can be used to encode different kinds of text. At the heart of the experiment was a Bob Dylan song and its associated video, which incorporated text. The encoded text was overlaid the video and posted to youtube and a blog with links to the TEI website. At the same time, baseline line data over a four month period was gathered to understand the current community of practice with TEI. Data on hits to the TEI website and video website were collected to determine the number of hits to each site, IP addresses, timing of hits and dispersion of hits.

Preliminary Findings

At the time of writing this proposal, final data analysis of benchmark and viral experience is being completed. The final report will outline TEI's community of practice and provide recommendations for other organizations exploring similar issues.

Preliminary data analysis of benchmark data suggests that TEI's community of practice is larger than anticipated. While TEI's 82 member organizations compromise the core of this community, an additional 6,000 users visit the website and many download information about the guidelines. Of this group, about 130 members make up 80% of the visits. A surprising portion of this group is from non-English speaking countries.

In terms of the viral experience, the youtube video was viewed over 4000 times and was briefly the top-watched video in Canada. A significant portion of individuals then clicked through to the TEI website for further information. Interestingly, several blogs also picked up the video and directed their readers onto the video and TEI.

The benefits to the Digital Humanities community will be several. First, the paper outlines a means by which digital resources projects can identify and understand their community of practice and create awareness of digital resources and tools. Second, it supports recommendations made by others which advocate the need for specific actions to reach digital resource users.

References

Cohen, D. J., & Rosenzweig, R. (2005). Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web. Available from http://www.chnm.gmu.edu/digitalhistory/audience/index.php.

Cook, S. (2008). The Contribution Revolution: Letting Volunteers Build Your Business. *Harvard Business Review*, 86(10), 60-69.

Dobele, A., Toleman, D., & Beverland, M. (2005). Controlled infection! Spreading the brand message through viral marketing. *Business Horizons*, 48(2), 143-149.

Freedman, D. H. (2007). Mitchell Baker and the Firefox Paradox. Inc. Retrieved from http://www.inc.com/magazine/20070201/features-firefox.html.

Guthrie, K., Griffiths, R., & Maron, N. (2008). *Sustainability and Revenue Models for Online Academic Re-*

sources: An Ithaka Report. New York, New York: Ithaka.

Smith, T., Coyle, J. R., Lightfoot, E., & Scott, A. (2007). Reconsidering Models of Influence: The Relationship between Consumer Social Networks and Word-of-Mouth Effectiveness. Journal of Advertising Research, 47(4), 387-397.

Warwick, C., Terras, M., Hunginton, 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, UK: University College London.

[1]Title thanks to the blog thesecretmirror.com

Animating the Knowledge Radio

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

Stéfan Sinclair
McMaster University
sgsinclair@gmail.com

0. Introduction

Most text analysis tools either pre-index texts or work on smaller corpora in order to give users an interactive environment where they can ask questions of texts. High Performance Computing facilities provide the opportunity to develop analytical tools that process large amounts of data in real time and that lets us animate analysis. This paper presents a hypothesis about how animated analysis can take advantage of HPC facilities to provide useful information to the user, especially when the results of an analytical process are complex visualizations. In this paper we will do three things:

- Discuss the outcomes of the April 2008 workshop on the Digital Humanities and High Performance Computing at McMaster University. In particular we will draw attention to the opportunities for digital humanists to develop a high resolution visualization agenda.
- Demonstrate the Big See model for large-scale text visualization and discuss animated visualizations.
- Describe a preliminary model for a streaming Knowledge Radio.

1. Outcomes of the Digital Humanities and HPC Workshop

In April 2008 SHARCNET, an HPC consortium in Southern Ontario, hosted a workshop on bridging the gap between digital humanities research and HPC facilities. The workshop brought together digital humanities researchers and HPC researchers from the SHARCNET participating universities to discuss the challenges and identify the opportunities for collaboration. The mission of the workshop was to ask and begin answering the following questions:

- What are the opportunities for the use of HPC facilities for humanities research?

- What examples are there of good practices and research innovation?

- What are the barriers to humanists using HPC facilities like SHARCNET and how can they be overcome?

- What concrete steps can SHARCNET (and by extension other HPC facilities) take to reach out to computing humanists and to then support their research?

In the spirit of open research the notes, action items, and report at available online (See https://www.sharcnet.ca/dh-hpc/index.php/Main_Page). Some of the key initiatives that were recommended to SHARCNET were:

- Humanists need introductory documentation about HPC and its uses in the humanities. John Bonnet, working with Geoffrey Rockwell and Kyle Kuchmey prepared an introduction to High Performance Computing in the Arts and Humanities with some examples. (See http://www.sharcnet.ca/Documents/HHPC/hpcdh.html)

- For humanists to take advantage of HPC facilities we need **training** opportunities, access to **fellowships** that place us into HPC facilities, and opportunities to **prototype** ideas with HPC folk (what we call charettes).

- The Digital Humanities and other disciplines that use HPC facilities should collaborate around information visualization. The arts and humanities have much to offer in the way of visual ideas and traditions of interpreting the visual. Humanists likewise can make good use of visualization facilities that are developed by HPC facilities.

One of the key opportunities identified were in large-scale and/or high resolution visualization. How would we represent evidence in the humanities if the size of the display, the resolution of the display and the speed of processing were not an issue? And that is the subject of this paper - experiments in developing a model for large-scale representation.

2. The Big See and Animated Visualization

The Big See is an experiment in high-performance text visualization. The project has developed a prototype of how a text or corpus of texts could be represented if processing and the resolution of the display were not an issue. Most text visualizations, like word clouds and distribution graphs, are designed for the personal computer screen. In the Big See we anticipate wall displays with 3 dimensional display capabilities and the processing to manipulate large amounts of data like all the content words of a corpus in real time. The Big See proposes one visual idea of what such a high performance visualization would look like as it is generated and once it is manipulable. To be clear, the current version of TBS does not need to run on an HPC system, it can run on a PC, but it models visual ideas for anticipated wall display systems. The question we asked ourselves was:

> *How could we represent a text or corpus if we had high resolution, wall-sized displays and processing was not an issue?*

The idea was to take a 2-dimensional visualization that was based loosely on the successful Weighted Centroid model that has been beautifully implemented by TextArc, and to add a third dimension to it and real time manipulation. The Big See in its default setting shows a pipe of the 20 most frequently used words, each of which is a line stretching the length of the pipe with markers where that word occurs. The pipe of lines can be turned in three dimensions so that it can be treated as a revolving barrel of distribution graphs. In the center of the pipe you have the text itself that recedes off into the distance much like the opening text of the Star Wars movies. Clicking on an instance of a word on the pipe advances the text to the appropriate point. The Big See is currently implemented as a PC application that has controls for the various parameters, though the source was written to be run on an HPC system.

Animated Visualization. One of the unanticipated outcomes of the project was that we found the live generation of the visualization compelling in and of itself. It has the virtue that it makes the final visualization understandable as you can see how, as the text is processed (and marches off into the horizon), the high frequency vocabulary changes. It also has the virtue that is animates a computer reading of the text in a linear fashion (starting with the first word and updating the visualization word by word. Users can infer things about the vocabulary of the text as it proceeds, though we have to be careful not to confuse the computer reading (and animating) a text with human reading, which is not necessarily so linear. This leads us to hypothesize that the animation of analytical processes can bear useful information for users of text analysis tools if properly paced and if they can represent the process. An animation can stand in for a pragmatic demonstration of what the computer is doing in its black box - “it reads in a word and adds it to the line for that work moving the line towards the 12:00 noon spot on the centroid if the frequency surpasses that of another word." Animations have, along with interactivity, the *prima facie* capability to bring processes alive giving users an intuitive understanding of what the final visualization represents. Obviously, they also have the ability to mislead the user, which suggests a fruitful avenue for further research. What are the best practices in information animation?

3. The Knowledge Radio

The Knowledge Radio is an extension of work done on the Big See, but adds the following aspects:

- Instead of working with a large, static corpus, what if we were to work with a large, open-ended or dynamic corpus where new input would modify the visualization as it was being processed (what Ben Fry calls organic information visualization)?
- What might be the most useful types of information to display to users for a dynamically analyzed diachronic corpus?

An example use-case for the Knowlege Radio is a blogger who wants to examine how a specific concept evolves over time on the web. The blogger would provide a search term or semantic field to a tool that would begin querying a search engine by successive date ranges and provide visual information on aggregate data as it was crawling results. The blogger could fine-tune the parameters of the visualization and scrub along a timeline to replay certain moments, or to fast-forward to the current point of analysis. The visualization might resemble something like the "code_swarm" project, which represents commit activity by individuals in several open-source project (http://vis.cs.ucdavis.edu/ ~ogawa/ codeswarm/).

From a tool implementation perspective an interesting challenge is to process a corpus as a stream rather than as a static object, especially in ways that would permit the user to playback segments, to compare segments, to modify parameters for visualizing segments, and to summon previously processed text for closer reading. This requires a well planned model for maintaining and updating aggregate data (as new text is processed) but also for storing relevant data from previously processed text.

This portion of the presentation will focus on describing the theoretical aspects and design principles of the Knowlege Radio but will also briefly demonstrate the state of the current Knowlege Radio prototype.

4. Links

Fry, Ben, "Organic Information Visualization" <benfry.com/organic/>

Ogawa, Michael, "code_swarm" <vis.cs.ucdavis.edu/~ogawa/codeswarm/>

Paley, Bradford, "TextArc" <www.textarc.org/>

Rockwell, Geoffrey et al. "The Big See" <tada.mcmaster.ca/view/Main/BigSee>

SHARCNET: <www.sharcnet.ca/>

“Going to the Show”: Spatial and Temporal History of Moviegoing in North Carolina

Natasha Smith
University of North Carolina at Chapel Hill
Libaries
nsmith@email.unc.edu

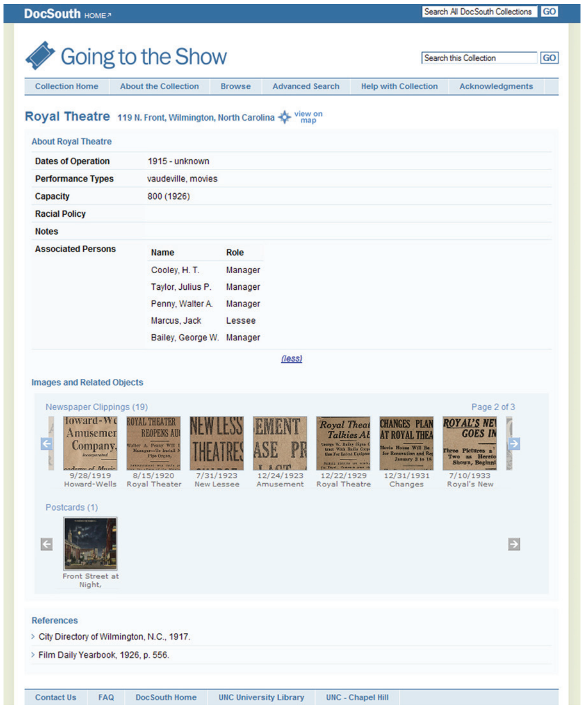
Elise Moore
University of North Carolina, Chapel Hill
elimoore@email.unc.edu

Kevin Eckhardt
University of North Carolina, Chapel Hill
kevineck@email.unc.edu

Robert C. Allen
University of North Carolina, Chapel Hill
rallen@email.unc.edu

A scholarly digital publication “Going to the Show” grew out of Prof. Robert C. Allen’s use of UNC Library archival materials in teaching and writing about the history of film exhibition and moviegoing. It is a result of close collaboration between Prof. Allen, James Logan Godfrey Distinguished Professor of American Studies, History, and Communication Studies at the University of North Carolina at Chapel Hill and *Documenting the American South* (DocSouth)—a digital library laboratory that creates, develops, and maintains online digital collections regarding the history of the American South drawn primarily from the outstanding archival holdings of the UNC library. “Going to the Show” documents and illuminates the experience of movies and moviegoing in North Carolina between the introduction of projected motion pictures (1896) and the end of the silent film era (1930). It is a historiographic experiment on several levels, in which collaborators endeavor to use digital technologies in a variety of innovative ways to collect, organize, and display data and materials that illuminate the historical experience of cinema.

As theatrical moviegoing becomes a thing more remembered than experienced, Allen argues that one of the most striking features of the experience of cinema for a hundred years was its sociality. For a century following the demonstration of Edison’s Vitascope projector at Koster and Bial’s Music Hall in New York on April 23, 1896, the experience of cinema in America and, indeed, around the world, involved a distinctive (though highly variable) social practice: groups of people converging upon particular places to experience together something understood to be cinema. The project team selected and published online numerous city directories, local newspapers ads and clippings, several hundred picture postcards and photographs that assisted with telling the story in a fuller or more holistic manner (see Figure 1.)



Royal Theatre 119 N. Front, Wilmington, North Carolina

About Royal Theatre

Dates of Operation	1915 - unknown
Performance Types	vaudeville, movies
Capacity	800 (1926)
Racial Policy	
Notes	
Associated Persons	

Name	Role
Cooley, H. T.	Manager
Taylor, Julius P.	Manager
Penny, Walter A.	Manager
Marcus, Jack	Lessee
Bailey, George W.	Manager

Images and Related Objects

Newspaper Clippings (19) Page 2 of 3

9/28/1919 Howard-Wells
8/15/1920 Royal Theater
7/31/1923 New Lessee
12/24/1923 Amusement
12/22/1929 Royal Theatre
12/31/1931 Changes
7/10/1933 Royal's New

Postcards (1)

Front Street at Night,

References

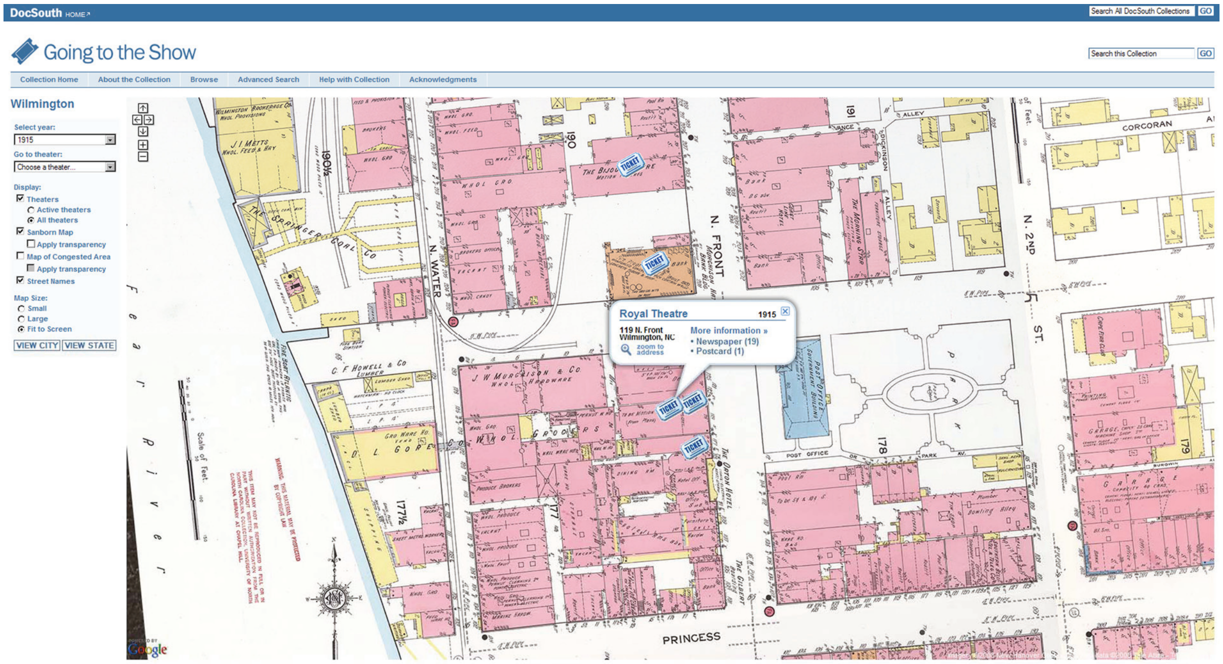
- City Directory of Wilmington, N.C., 1917.
- Film Daily Yearbook, 1926, p. 556.

Figure 1. Metadata associated with Royal Theater in Wilmington, NC. Additional resources – postcards, newspaper ads and clippings, city directories.

However, recent work in cultural geography has helped us to look at the history of cinema *as* the history of the experience of cinema and to look at cinema as a set of practices occurring in space (see Doreen Massey). Like space, the experience of cinema is relational, heterogeneous, and open-ended. In that, collaborators turned to another valuable resource of information–Sanborn Fire Insurance Maps. Between 1867 and 1977 the Sanborn Map Company of Pelham, New York, produced large-scale (usually 50 feet to the inch) color maps of commercial and industrial districts of some 12,000 towns and cities in North America to assist fire insurance companies in setting rates and terms. Each set of maps represented each built structure in those districts, its use, dimensions, height, building material, and other relevant features (fire alarms, water mains and hydrants, for example).

Looking through thousands of map pages over the years, thinking about how they represent the experience of moviegoing and about how the maps might be represented in “Going to the Show,” Allen started rethinking not only the sociality but also the spatiality of the experience of cinema more generally. For most white people living in towns or cities of any size in N.C. and those living in the countryside around these towns and cities in the first three decades of cinema history, going to the movies was a part of the experience of the spaces of downtown social, cultural, commercial and consumer life. Understanding what went on inside the theater requires understanding what went outside.

What the Sanborn maps enable us to see—in ways that other representations of the social experience of moviegoing do not—is that the space of the experience of cinema in towns and cities across North Carolina, and possibly in many other places as well, was not bounded by the places in which movies were shown. The maps show clearly that the emergence of movie culture in North Carolina is inextricably linked to the rise and development of urban central business districts (see Figure 2.)



Wilmington

Select year: 1915

Go to theater: Choose a theater...

Display:
☑ Theaters
○ Active theaters
● All theaters
☑ Sanborn Map
□ Apply transparency
□ Map of Congested Area
□ Apply transparency
☑ Street Names

Map Size:
○ Small
○ Large
● Fit to Screen

Royal Theatre 1915
119 N. Front
Wilmington, NC
zoom to address
More information »
• Newspaper (19)
• Postcard (1)

Figure 2. Wilmington, N.C., 1915. Hotspot “TICKET” displays Royal Theater with relevant available information.

We will represent the Sanborn maps in “Going to the Show” in a way that they were never intended to be: with individual map pages digitally stitched together so that they form a composite overview of a town’s central business district. In 45 communities, moviegoing is mapped onto Sanborn Fire Insurance Maps (1896-1922), showing how moviegoing became an important part of the experience of everyday life in the growing cities and towns of early 20th century North Carolina. These map pages have been digitally “stitched” together and geo-referenced to display the entire “downtown” of representative communities across the state. The evolution of successive map sets produced from 1896 to 1922 illustrates how these towns and cities grew and changed at a time of rapid urbanization and industrialization. Displaying the maps in Google Earth vividly illustrates a century of continuity and change.

The decision process for which tools to use in order to bring this picture to life, has been a balance between experimental technology and usability. “Going to the Show” is using kml layers upon Google Maps and Google Earth to place the Sanborn maps within context. Google’s open-source map API is used for zooming and hotspot addition. There is also a MySQL (soon to be PostGRES) database to manage the metadata about the contextual items which have been collected in order to paint the entire picture of moviegoing. As a result, users can search for movie exhibition sites, locations, managers, and racial policy across the first thirty-five years of moviegoing in more than 200 N.C. communities. Adobe PhotoShop is used for stitching together digital Sanborn Map images and basic manipulation of documentary content images, as necessary (sharpening, resizing, and more). The software Global Mapper is used for geo-referencing Sanborn maps. It converts, edits, prints, tracks GPS, and allows utilization of GIS functionality in datasets.

References

Richard Maltby, Melvyn Stokes, and Robert C. Allen, eds., **Going to the Movies: Hollywood and the Social Experience of Cinema** (Exeter: Univ. of Exeter Press, 2007)

Doreen Massey, **For Space** (London: Sage, 2005)

Anne Kelly Knowles, **Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship** (California: ESRI Press, 2008)

Suzy Beemer**,** Richard Marciano**,** Todd Presner, **Seeing Urban Spaces Anew at the University of California** (CTWatch Quarterly, May 2007, vol. 3, number 2, http://www.ctwatch.org/quarterly/articles/2007/05/seeing-urban-spaces-anew/1/)

What is transcription? (part 2)

C. M. Sperberg-McQueen
World Wide Web Consortium
cmsmcq@acm.org

Claus Huitfeldt
University of Bergen
Claus.Huitfeldt@fof.uib.no

Yves Marcoux
Université de Montréal
Yves.MARCOUX@UMontreal.CA

In earlier work, Huitfeldt and Sperberg-McQueen have sketched out a series of formal models of transcription, characterized by gradually increasing complexity and descriptive power, but not, in the end, by a satisfactory explication of transcription [Huitfeldt/Sperberg-McQueen 2008].

The essential proposition of these models may be called the Reading Identity Hypothesis: namely, that in simple cases the relation between a transcription and its exemplar is that each can be read as containing (a sequence of tokens which map to) the same sequence of types. When types and tokens are identified at the character level, for example, a document containing marks which can be read as tokens representing the sequence of characters

> Come, sit thee downe upon this flowry bed,

can be regarded as a transcription of another document just in case that document, too, contains marks which can similarly be read, as tokens of those types, in that order. The addition of various meta-characters in transcription (e.g. explicit marking of line breaks in the exemplar) is modeled by allowing readings to be selective, ignoring some marks in the document being read. Some common variation in transcription practice is modeled by postulating different token/type mappings with different type inventories: the example just given, for example, will count as a transcription of one line in Shakespeare's First Folio only if long and short s are treated as the same grapheme and if u and v are normalized. When long and short s are not mapped to the same grapheme, and u and v are distinguished based on their letter form and not on current usage, then a transcription will contain a sequence of tokens readable as:

Come, ſit thee downe vpon this flowry bed,

Prominent among the shortcomings of the reading-identity hypothesis as formulated in [Huitfeldt/Sperberg-McQueen 2008] is the explicit assumption that a document, and its transcription, consist essentially of a sequence of characters (or more generally of tokens).

In this paper, we will assume that a document identifies not just a sequence of typed tokens, but a structure consisting of typed relations (such as containment or dominance) between sequences or sets of typed objects (such as paragraphs or section headings). The relation between documents can then be modeled as a relation between such structures, allowing transcriptions to agree or disagree not only on character sequences, but also on document structure, and on assignments of types to document objects and their relations.[1]

Based on these assumptions we will propose a formal model of document structure which is able to capture basic notions underlying conventional markup formalisms such as the tree structures of XML, the concurrent trees of XCONCUR [Hilbert et al. 2005], the sequence-plus-ranges model of LMNL [Tennison/Piez 2002], and the directed acyclic graphs of Goddag structures [Sperberg-McQueen/Huitfeldt 2000]. The formalism we develop is not intended and should not be understood as an alternative to any of these, only as a way of trying to reduce them to a common denominator in terms of which it is possible to determine that two transcriptions do or do not agree on particular aspects of the exemplar.

We believe the model can be used to explain, at an appropriate level of abstraction, how it is possible that representations using different markup formalisms and their associated vocabularies can agree or disagree on the transcription of a given exemplar.

For example, in most current practice, the line given above as an example is likely to be embedded in a document structure showing that it occurs within a particular play, act, scene, and speech: for example

```
<play>
  ...
  <act>
    ...
    <scene>
      <speech>
        <speaker>Tita.</speaker>
        <l>Come, sit thee downe vpon this
flowry bed,</l>
        ...
      </speech>
      ...
    </scene>
    ...
  </act>
  ...
</play>
```

or, using a different XML vocabulary (TEI),

```
<text>
  ...
  <body>
    ...
    <div type="act" n="IV">
      <head>Actus Quartus</head>
      <div type="scene" n="i">
        <sp who="Titania">
          <speaker>Tita.</speaker>
          <l>Come, &longs;it thee downe &v;pon
this flowry bed,</l>
          ...
        </sp>
        ...
      </div>
      ..
    </div>
    ...
  </body>
  ...
</text>
```

In older electronic texts the transcription is likely to use a different markup language entirely (here COCOA markup):

```
<T Midsummer Night's Dream>
<A 4> ...
<S 1>
<C Titania>
Come, *sit thee downe vpon this flowry bed,
...
```

These three transcriptions identify, by means of their markup, textual structures which are substantially similar, but different in details. All three, for example, can be read as indicating that the line in question is part of the first scene of the fourth act, and that it is spoken by Titania, but they differ in details (act 4 or Act IV or Actus Quartus? “Tita.” or “Titania”? or both?)

One challenge for a formal model of transcription is to distinguish essential differences among these transcriptions, which represent disagreements about the text of the exemplar (and thus disagreements among the transcribers) on the one hand, from inessential differences (or in some cases essential differences which happen not to relate to questions of transcription per se) which follow from the different markup languages used, or from the choice of textual features to mark up, on the other hand.

The main body of the paper is the definition of the formal model, followed by an extended example showing the transcription of a short document in several markup languages and the representation or analysis of these transcriptions in the abstract model of transcription.

The formal model associates to each transcription a set of statements (or assertions) about the transcribed exemplar. The statements may be expressed in any convenient notation, e.g. some form of first-order logic. Since different transcription practices and contexts may consider widely different sets of textual features as relevant, the exact predicates allowed in the statements can vary from one transcription to another.

Essentially, the statements corresponding to a transcription are obtained as follows. First a set of basic facts about the exemplar is derived in a straightforward manner directly from the encoding of the transcription. Then, the closure of that set under inference is taken. The inference rules will vary with the specific transcription practice, conventions, and context in which the transcription was performed, and can also include special rules for "translating" statements from one combination of transcription practice/conventions/context to another.

If two transcriptions agree in every detail (regardless of how they are coded), the sets of inferences derived from them will be equal. Intuitively, if one set is a subset of the other, this indicates that the transcriptions do not disagree, but that one of them is more detailed or thorough than the other.

If neither set is a subset of the other, then we have to look at the (closure under inference of the) union of the sets. If the union is consistent (i.e., does not allow inferring both a statement and its negation), the transcriptions still do not disagree, but each include aspects of the exemplar that the other does not. However if the union is inconsistent, then we can conclude that the transcriptions disagree, and even say on what particular point(s) they disagree.

Note that the construction of the set of basic facts from the encoding of the transcription is not a simple mapping from generic IDs (or, in general, markup configurations) to predicate names. Thus, for instance, different generic IDs could map to the same predicate name. Some markup may also be ignored, and even entire elements, like elements that are not part of the transcription as such (for example, the TEI header[2]). More complex methods of constructing the set of basic facts are also possible.

Apart from its intrinsic interest, a formal model of transcription is of use in efforts to support a formal description of the meaning of various constructs in well known markup schemes, in which the meaning of certain elements is inextricably linked to the fact that they can contain, or be used in, transcriptions of existing documents. We hope to conclude with some observations about the applicability of our model in the formal description of markup vocabularies.

References

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. "What is Text, Really?" *Journal of Computing in Higher Education* 1.2: 3-26.

Hilbert, Mirco, Oliver Schonefeld, and Andreas Witt. 2005. "Making CONCUR work." In *Proceedings of Extreme Markup Languages 2005*. On the Web at <URL:http://www.idealliance.org/papers/extreme/proceedings/html/2005/Witt01/EML2005Witt01.xml>

Huitfeldt, Claus, and C. M. Sperberg-McQueen. 2008. "What is transcription?" L&LC 23.3 (2008): 295-310. Available on the Web at <URL:http://llc.oxfordjournals.org/cgi/reprint/fqn013?ijkey=97G3O9T2QOGozEm&keytype=ref>

Power, Richard, Donia Scott, and Nadjet Bouayad-Agha. 2003. "Document structure". *Computational Linguistics* 29.2: 211-260.

Renear, Allen, David Durand, and Elli Mylonas. 1996. "Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies". In *Research in Humanities Computing* ed. Nancy Ide and Susan Hockey. Oxford: OUP, 1996.

Sperberg-McQueen, C. M., and Claus Huitfeldt. 2000."GODDAG: A Data Structure for Overlapping Hierarchies" paper given at *Digital Documents: Systems and Principles. 8th International Conference on Digital Documents and Electronic Publishing*, DDEP 2000, *5th International Workshop on the Principles of Digital Document Processing*, PODDP 2000, Munich, Germany, September 13-15, 2000. Published in DDEP-PODDP 2000, ed. P. King and E.V. Munson. Lecture Notes in Computer Science 2023. Berlin: Springer, 2004, pp. 139-160. Available on the Web at <URL:http://www.w3.org/People/cmsmcq/2000/poddp2000.html>

Tennison, Jeni, and Wendell Piez. 2002."The layered markup and annotation language (LMNL)"Late-breaking talk given at *Extreme Markup Languages 2002* (not in proceedings). Slides on the Web at <URL:http://www.idealliance.org/papers/extreme/proceedings/author-pkg/2002/Tennison02/EML2002Tennison02.zip>

Notes

[1] This assumption in some ways resembles the celebrated assertion, widely known as the OHCO hypothesis, that "text is an ordered hierarchy of content objects" [DeRose et al. 1990] but differs from it in some ways. We are speaking about documents (physical objects) not texts (abstract objects); we say that documents identify certain abstract structures not that they are constituted by those structures; our assumption does not imply that the abstractions involved are "content objects", nor that they are hierarchically structured, nor that they are ordered. Our assumption does however share with the OHCO hypothesis the idea that documents have structure more complex than a simple sequence of types or tokens. See also the retraction of the hierarchical part of the OHCO hypothesis by several of its authors [Renear et al. 1996] and the linguistically motivated analysis of document structure in [Power / Scott / Bouayad-Agha 2003].

[2] Elements such as the TEI header may however be very helpful in determining the practice/conventions/ context of the transcription.

Our Americas Archives Partnership: Charting New Cultural Geographies

Lisa Spiro
Rice University
lspiro@rice.edu

Over the last decade, Hemispheric American Studies has been an emerging interdisciplinary field, spawning new conferences, graduate programs, professional seminars, journals, essay collections, and even a scholarly association. As Caroline Levander (co-PI of the Our Americas Partnership) and Robert Levine explain, Hemispheric American Studies aims "to chart new literary and cultural geographies by decentering the U.S. nation and excavating the intricate and complex politics, histories, and discourses of spatial encounter that occur throughout the hemisphere but tend to be obscured in U.S. nation-based inquiries" (Levander & Levine 2008, p.5). Such an approach means being attentive to shifting borders, migrations, cultural interactions—movement rather than fixity.

Scholars practicing a hemispheric approach to American Studies face challenges that include research materials being scattered in many repositories, lack of institutional support and intellectual community at their home institutions, and linguistic barriers. The Our America Archive Partnership (OAAP, http://oaap.rice.edu/) is beginning to address these limitations by building a distributed online collection of research materials as well as technologies for federating multiple repositories, creating geographical and temporal visualizations, and enabling researchers to tag and collect digital objects. Funded by a $1 million grant from the Institute for Museum and Library Services (IMLS), OAAP brings together Rice University's Fondren Library, Rice's Humanities Research Center, the University of Maryland's Institute for Technology and the Humanities (MITH), and Mexico's Instituto Mora, with more partners expected. Included in the OAAP project are evaluation of scholarly needs; digitization of texts; development of an open federated archive; and design and implementation of technologies that support inquiry, discovery and collaboration.

The Archive

With its transnational scope, Hemispheric American Studies requires access to research materials distributed around the world, with collections that transcend nation state boundaries and facilitate comparativist studies of the Americas. The web-based OAAP provides such access, with tools that enable scholars to search for resources, collect and organize them, visualize them using temporal and geospatial interfaces, and collaborate with peers. OAAP brings together an existing collection of TEI-encoded texts, the University of Maryland's Early Americas Digital Archive (EADA, which covers literature of the Americas from 1492-1820), with two new collections being digitized, Rice's America's Collection and Instituto Mora's collection (North and Latin American government documents, manuscripts, books, and other material from 1811-1920). The combined collections span the five hundred year period that witnessed the making of colonial and modern cultures in the Americas. Because of its range, the federated archive promises to reinvigorate the study of American literary and cultural history by suggesting unexpected juxtapositions, different models of periodization, and new avenues of cross-cultural influence.[1]

User Studies and Needs Analysis.

The OAAP team is working actively with the Hemispheric American Studies scholarly community in assessing needs, building the collection and tools, and developing an infrastructure for collaboration. Caroline Levander, a leading researcher in Hemispheric American Studies, is a co-PI, and Rice's Humanities Research Center, which Levander directs, contributes most of the scholarly activities for the project. The advisory board includes well-respected scholars, librarians, and digital humanists. Levander has taught an NEH Summer Seminar on the topic of hemispheric American literature and a National Humanities Center Dupont seminar on the globalization of American literary studies. Participants were surveyed about their existing use of digital resources as well as about the OAAP and will be consulted as OAAP continues to evolve. We also collaborated with the Council on Library and Information Resources to convene a meeting of five scholars and one academic technologist to envision OAAP's future.

From these surveys and discussions several principles have emerged:

- *Access is paramount.*

  Seminar survey respondents ranked better search tools and better access to research materials as top priorities. Researchers emphasized the importance of having access to a variety of materials from around the world, from government documents to manuscript materials to audio recordings. These resources must be easy to discover through full-text search as well as rich metadata.

- *OAAP should tackle barriers of language and discipline*

  How can scholars get beyond the limitations of their own conceptual schemas and their own languages? One scholar envisioned a search tool that would work like a friendly archivist, identifying materials you didn't know existed. Scholars emphasized the need for a multilingual search interface, so that a researcher could enter keywords in Spanish and pull up relevant results in English.

- *Scholars want an environment that facilitates collaboration, acknowledges contributors, and enables them to quickly evaluate the reliability of resources.*

  Fundamentally scholars saw OAAP as opening a new approach to scholarship, one that is not constrained to close readings of a few key texts but able to encompass broad spatial, political, and social contexts. For instance, a group of scholars could team up on a study of the Spanish-American War, some focusing on visual or literary culture, some on demography. One survey respondent suggested that OAAP "develop [a] place for scholars to ask one another questions about materials with which they are less familiar, i.e. provide space for collaborative interactions/project."

Text Encoding Approach

Texts from OAAP's collections are marked up in TEI to facilitate scholarly analysis and long-term preservation. Since scholars seem to value access over all else, some have suggested that it would be sufficient to offer PDFs of the books. With TEI marked-up texts, however, scholars can search within or across texts, use analytical tools such as TAPOR, and easily copy out passages into their notes. Moreover, TEI supports the creation of multiple forms of output from a single TEI text, such as an original and a regularized version of a manuscript or a version optimized for mobile devices.

After consultation with scholars, OAAP decided to focus on marking up basic structures in the body of the text, just as EADA did with its TEI texts. Our aim is to support access and analysis, not to create critical editions. For manuscripts, we are also encoding features such as deletions, additions, and corrections, as well as the structure of letters such as openers and closers.

Given the significance of geography to OAAP, we considered using automated or manual methods to add place markup. While this could enable geospatial searching or the generation of maps showing places described in a text, it would not describe the complexities of place, such as the distinctions between places in which a narrator is physically present vs. imagined spaces (like "home"). We needed a streamlined approach that would minimize labor costs in working with the 25,000 pages of digitized text. Furthermore, to encourage other institutions to add their collections to the federation, we want to keep the requirements for participating in OAAP minimal. We must, therefore, pursue a balanced approach that embraces standards (to facilitate consistency and longevity) and is flexible enough to accommodate a diversity of archives with a range of goals and capabilities.

Federation Model and User Interface

In addition to providing access to key research materials, OAAP hopes to facilitate collaboration among an international community of scholars. Despite challenges, including the humanities' reluctance to reward collaborative work, scholars in this vital community have expressed excitement about working together on research projects in the collaborative space provided by the OAAP. The project is facilitating collaboration through a social tagging interface and federation of distributed repositories, as well as by working actively with scholars on the OAAP's development. To enable the federation of distributed repositories, OAAP is implementing a harvester that collects objects from repositories meeting minimal standards and stores them for the purposes of searching and display. OAAP is adopting a data exchange standard based upon OAI's Static Repository XML model (http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm). Through OAAP's interface, researchers will be able to search, browse, tag, and collect objects from distributed collections. Researchers will tag objects to create a set of keywords that others can use, thus expanding the metadata beyond the Library of Congress Subject Headings added by metadata librarians and OAAP keywords provided by participating scholars. OAAP will also implement faceted browsing, allowing researchers to filter searches by, for instance, "political constitution" and "1800-1850." Since notions of time and space are crucial to OAAP, it will provide both geospatial and temporal interfaces, enabling researchers to view resources on an interactive map or timeline.

Conclusion

By providing open access to core research materials, federating archives, and developing tools for analysis and collaboration, OAAP hopes to realize what Cathy Davidson calls Humanities 2.0: "Hybridity, exchange, flow, and cultural transaction are all explored more responsibly and adventurously when the resources of many nations, in many languages, have been digitized, made interoperable, and offered for research by scholars around the world, each of whom brings a local store of knowledge and experience to the theoretical, interpretive enterprise" (Davidson 2008).

References

Davidson, C.N., 2008. Humanities 2.0: Promise, Perils, Predictions. *PMLA*, 123(3), 707-717. Available at: http://www.mlajournals.org/doi/abs/10.1632/pmla.2008.123.3.707 [Accessed October 5, 2008].

Levander, C.F. & Levine, R.S., 2008. Introduction. *Hemispheric American studies*, New Brunswick, N.J.: Rutgers University Press, pp. 1-17.

[1] See *Our Americas Archive Partnership Narrative*, http://oaap.rice.edu/documentation.html

Sentiment Analysis of Fictional Characters Based on Entity Profiles

Rohini K. Srihari
University at Buffalo
rohini@cedar.buffalo.edu

Laurie Crist
Janya, Inc
lcrist@janyainc.com

Harish Srinivasan
Janya, Inc
hsrinivasan@janyainc.com

Introduction

There has been an increasing interest in applying automatic text analysis techniques to various text classification problems in literature and the social sciences. Examples of such tasks include determining biases in political coverage, or analyzing mood in literature. Earlier techniques were based on simplistic corpus analysis techniques such as counting word frequencies and co-occurrences. The access to robust machine learning technology and tools has enabled more sophisticated text mining techniques to be developed. Yu (Yu, 2008) discusses the use of text classification methods in the literary domain. His study compared the performance of two popular algorithms, naïve Bayes and support vector machines (SVMs) in two literary text classification tasks. While this trend represents progress in automatic text mining, it still reflects a reliance on primitive features such as the *bag-of-words* model. In such models, text is represented as a vector of weighted words; word order is disregarded and only frequency information is used. Such techniques are inherently limited in the granularity of the analysis they can perform, typically limited to the document level. For more fine-grained tasks such as sentiment analysis with respect to people, characters or topics, a more sophisticated model of relevant context is required.

This work discusses the use of *entity profiles* to represent the context in which to make judgments regarding an entity, where an entity can represent an individual or an organization, or other salient entity types. An entity profile reflects a consolidation of all important information pertaining to an entity within a document. For a person (or a character in a novel), the entity profile would include all mentions of the individual, including co-referential mentions, as well as relationships and events involving the person. The representation of such information is typically highly structured such as `spouse_of(Maria Bertram, Mr. Rushworth)` with a link to the *text* snippet or sentence from which the relationship or event was extracted. An entity profile, when compiled from a collection of documents, or a lengthy novel is rich information that provides the required context in which to compare two individuals, classify human behaviour, etc. Automatically extracting entity profiles (and associated text snippets) is a challenging task in information extraction; the next section describes a system which has been designed for this purpose. The rest of the paper describes the use of entity profiles as the context in which automatic sentiment analysis (Chesley et al, 2006) of fictional characters can be computed. The example is from Jane Austen's Mansfield Park.

Semantex: An Information Extraction Engine

Semantex (Srihari 2008) is a domain independent, intermediate level information extraction (IE) engine. The linguistic processor modules support different levels of natural language processing, including orthography, morphology, syntax, co-reference resolution, semantics, and discourse. The categories of information objects created by Semantex are (i) Named Entities (NE): proper names of persons, organizations, product, location etc., (ii) Correlated Entity (CE) relationships: capture local relationships between entities within sentence boundaries. The results are consolidated into EPs based on co-reference and alias support, (iii) Entity Profiles (EP): Entity Profiles are complex rich information objects that collect entity-centric information—in particular, all the individual mentions of an entity in a document and any CE relationships the entity is involved in, (iv) Subject-Verb-Object (SVO) triples: SVO triples decoded by Semantex are logical, rather than syntactic: surface variations such as active voice vs. passive voice are decoded into the same underlying logical relationships, (v) General Events (GE): verb-centric information objects representing `who did what to whom when and where'. These five types of information objects capture key content of the processed text. For this project, the most relevant objects are CEs, EPs, and SVOs.

Sentiment Analysis based on Context provided by Entity Profiles

We use the set of text snippets (or sentences) from an entity profile as the context in which features for sentiment analysis are computed. Sentiment analysis is performed in two phases: (i) the first phase, training, focuses on compiling a lexicon of subjective words and phrases along with their polarities (positive/negative) and an associated weight. (ii) in the second phase, sentiment association, a text document collection is processed and sentiment assigned to entity profiles of interest.

For sentiment analysis, a lexicon of subjective words/phrases (those with positive or negative polarity associated with them) is first compiled through (i) expansion from adjectives in WordNet using synonyms based on positive and negative seed adjectives and (ii) use of a search engine to find words that appear “near” a known positive/negative adjective. To associate sentiment with an entity, we accumulate polarity weights (using a sliding window) from the sentences within the entity profile; thresholding results in a final positive, negative or neutral polarity for the entity in question.

Sentiment Analysis applied to Jane Austen’s *Mansfield Park*

In this section, sentiment analysis has been applied to characters in *Mansfield Park* by Jane Austen. Specifically, it has been applied to the entity profile for the character Mary Crawford at different times in the novel. This is the process that was employed.

1. The text of Mansfield Park , originally consisting of 159,500 words was split into four parts at chapter breaks with some consideration to the progress of the plot. These breaks were chosen to track the transformation of the character Mary Crawford from first meeting through the revelation of some flaws in the character.

2. Each of the four sections was processed by Semantex; entity profiles were generated for all the characters, including Mary Crawford. This resulted in four entity profiles for Mary Crawford at different stages in the plot.

3. Sentiment analysis was computed for each of the entity profiles: the goal was to correlate the output of automatic sentiment analysis with the transformation in the character over time.

The sentiment analysis output based on two entity profiles for Mary Crawford generated at different stages (parts one and three) is shown in the table below. Part three reflects the duration of time just before and after Maria’s elopement with Henry Crawford, Mary’s brother. Mary’s reaction to this event exposes flaws in her nature, and contributes to a reader’s judgment of her character as negative. In each case, a subset of the subjective words that contribute to the overall polarity (positive or negative) are shown, along with snippets of text (based on entity profile) in which those words appeared. These text snippets are a subset of the sentences which contribute to the entity profile for Mary Crawford. The entire profile is not shown for space considerations. It should be noted that snippets from the entity profile are not necessarily contiguous.

Our system has judged the first profile to be positive, but the second one to be neutral rather than negative. This could be partly due to an aggregation of sentiment that is performed over the entire section. There is considerable effort that remains in improving the accuracy of automated sentiment analysis of fictional characters. For example, words such as “ashamed” and “embarrassed” are not necessarily associated with negative sentiment depending on the context. Another problem is proper association of the sentiment with the character in question. We continue to work on these issues.

Co-referential Mentions: Mary, Mary Crawford, Miss Crawford, she, herself, his sister

Sentiment	Part 1 of 4	Part 3 of 4
Polarity	Positive	Neutral
Probability	0.673	0.883
Subjective Words	agreeable 0.49 bad -0.601 beauty 0.98 clever 0.49 comfort 0.4 dearest 1.154 delighted 0.53 disinclination -0.506 displeased -0.49 elegance 0.49 happiness 0.49 inconceivable -0.49 laughingly 0.495 liking 0.446 lively 0.3 not mean 0.4 pleasant 0.5 pretty 0.4 proud 0.6 very clever 0.8	abominable -0.601 agitation -0.939 ashamed -1.96 attractive 0.6 be satisfied 0.586 betraying -0.751 bitter -0.601 comfort 1.6 dearest 2.308 discontented -0.651 embarrassed -0.651 frown -0.55 manners 2.003 nobly 0.92 satisfactorily 0.6 shame -1.069 smile 1.47 unpleasant -0.683 vain -0.865 vexation -0.54

Snippets	Mary Crawford was remarkably pretty... the manners of both were lively and pleasant... Mary was her dearest object...You will be kinder than Mary. Dr. Grant laughingly congratulated Miss Crawford... Miss Crawford ‘s beauty did her no disservice...Mary was satisfied with the Parsonage...and not at all displeased either at her sister...we find comfort somewhere	To Mary it was every way painful...and I wish they may be heartily ashamed of their own abominable neglect and unkindness...she thought there was a smile -- which made her blush and feel wretched...if he really loved her , and were unhappy too !. There was comfort...confusion to form the clearest judgment of Miss Crawford‘s meaning

Conclusions

This paper has described an experiment in which automatic sentiment analysis is used to illustrate either the change in a character, or the perception of the character by other characters over the progression of a story. Entity profiles provide rich context in which to attempt other tasks, such as measuring the similarity of characters, both within a novel, as well as across novels. Standard document similarity measures may be used to accomplish this.

The challenge to making this technique more robust is the accuracy of coreference, including anaphora resolution. Mistakes in this module can cause irrelevant sentences to be pulled into the entity profile, thus rendering the analysis inaccurate. Efforts are underway to improve this accuracy. Sentiment analysis can also be improved by fine tuning the association of subjective words with the correct character. Nevertheless, this is a more sophisticated method of performing text analysis with respect to analyzing human behaviour.

References

R. K. Srihari, W. Li, C. Niu and T. Cornell (2008) "InfoXtract: A Customizable Intermediate Level Information Extraction Engine," Journal of Natural Language Engineering, Cambridge U. Press, 14(1), 2008, pp.33-69.

P. Chesley, B. Vincent, L. Xu, and R. K. Srihari (2006) "Using Verbs and Adjectives to Automatically Classify Blog Sentiment”, Proc. AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, Stanford University, CA March 2006, AAAI Press, TR SS-06-03, pp.27-29.

Bei Yu (2008) An evaluation of text classification methods for literary study, Linguist Computing 23: 327-343.

Computer-Aided Palaeography, Present and Future

Peter Stokes
University of Cambridge
pas53@cam.ac.uk.

The field of digital palaeography has received increasing attention in recent years (Ciula 2005; Bulacu and Schomaker 2007; Stokes 2007/8; Rehbein *et al*. forthcoming). However, this subdiscipline is not generally accepted by the wider humanities community, and indeed some have argued that handwriting is inherently unquantifiable and cannot be analysed by digital means (Costamagna *et al*. 1995–96). This paper therefore asks what problems might be amenable to digital analysis and why existing methods have not been widely accepted, and then introduces new software designed to address these issues.

Present Issues in 'Traditional' Palaeography

One of the principle difficulties in palaeography is subjectivity, since palaeographers often express qualitative opinions rather than objective arguments and this can lead to an 'authoritarian discipline' which depends on 'the authority of the author and the faith of the reader' (Derolez 2003). Attempts to make the field more scientific began well before the use of computers, including an attempt to measure hundreds of letters by hand (Gilissen 1973), but these were laborious and problematic and the approach was largely forgotten. Others have drawn on forensic document analysis, since forensic document analysts have had to demonstrate the efficacy and objectivity of their methods in court and are therefore far ahead of palaeographers in this respect (Srihari *et al*. 2002; Davis 2007). Nevertheless, the differences between modern and medieval handwriting are significant. Amongst other things, medieval scribes were (usually) highly trained and can show very little variety from one scribe to the next, unlike modern writers; forensic analysts can generally collect a large corpus of known handwriting whereas palaeographers often have few or no known examples; and palaeographers cannot verify their results with the certainty that forensic analysts can often achieve.

In addition to these problems, palaeographical study has also been hampered by the lack of an established terminology, and this makes commmunicating arguments very difficult indeed. The *Comité international de paléographie latine* (CIPL) was founded in 1953 partly for this reason but, despite sixteen international meetings to date, it is nowhere near producing an accepted solution (but see Muzerelle 2003). However, the problem is receiving more attention with the increasing use of databases and online catalogues. Attempts to computerise existing catalogues have revealed inconsistencies in earlier sources, and the absence of accepted terminology has made digital resources difficult to search and almost impossible to interconnect.

Another issue is the volume of data to be analysed. Palaeographers have tended to consider small corpora of scribal hands, but these are not necessarily representative of wider practices, and any conclusions based on such samples are necessarily limited. Larger corpora are beyond traditional methods, however, since they can include hundreds of scribal hands with potentially thousands or tens of thousands of features (Stokes 2007/8). Digital methods can help, as databases and data-mining can both be used to manage large quantities of material. Similarly, if catalogues and other online resources followed standards in describing handwriting then data can be pooled from many different projects. However, as noted above, such standards do not yet exist.

Present Issues in 'Digital' Palaeography

These difficulties can all be addressed to some extent by digital methods, and new studies in 'digital palaeography' are going some way towards doing that (Ciula 2005; Bulacu and Schomaker 2007; Stokes 2007/8). However, promising as these seem, they have received almost no acceptance and relatively little interest from 'traditional' palaeographers. This is partly because the technology is not yet mature, and perhaps also because the attempts to date have generally involved small projects without the sustained funding or larger interdisciplinary groups that digital humanities often require (Pierazzo 2008). However, neither of these seems to be a complete explanation, and another problem may be one of understanding and engagement. Even if procedures are communicated clearly using recognised terminology, scholars still require a good understanding of many complex fields to fully appreciate and engage with the material, including post-graduate level probability theory, digital signal processing, computer graphics, and more, all skills which cannot be expected from palaeographers. This means that they cannot engage meaningfully with the results of digital methods and that we have replaced the human authority with a digital one.

Another difficulty is that all results should be reproducible or at least verifiable if any claim to objectivity is made. However, this is rarely done in practice. For example, most techniques for enhancing images require a great deal of *ad hoc* human intervention which cannot be accurately documented using most existing tools. Many scholars in the humanities use Adobe Photoshop for enhancing images (Craig-McFeely and Lock 2006), but Photoshop provides no easy way of recording actions such as colour-selection, nor is it always clear exactly how the proprietary algorithms work. Photoshop is undoubtedly useful and should not be discarded lightly, but the methods and tools must allow full documentation and reproducibility.

The lack of standard terminology for handwriting not only makes communication between scholars difficult, but it also makes databases of script all but impossible. Some less conventional approaches have been tried, such as the Manchester 'C11' database (Rumble 2005; Scragg *et al*. 2005), but this particular classification and interface is very opaque and subjective, so that neither search criteria nor result-sets are intelligible to users. Alternatives such as descriptive rather than visual criteria for searches have also been attempted (Stokes 2007/8), but these still depend on users understanding a specific and *ad hoc* vocabulary and so are difficult to search and are nearly impossible to integrate with other resources.

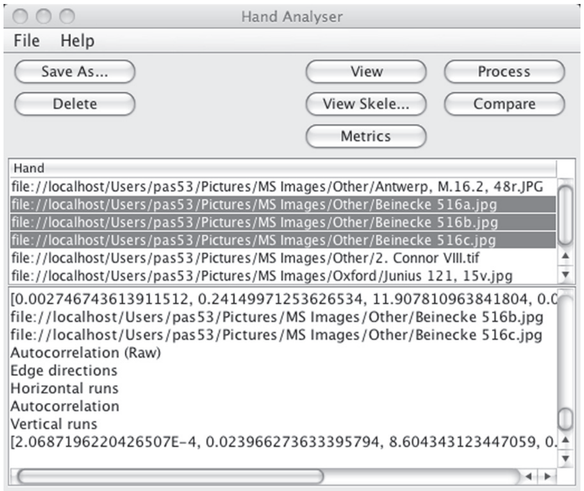
Some Future Directions

Perhaps the easiest problem to solve is that of documentation and reproducibility, since software can be designed to record all operations performed and then save this as a MIX/METS file or other standard format (Stokes, forthcoming). A Photoshop plugin could achieve this but would be subject to Adobe's licensing and the restricted availability of their Software Development Kit. Instead, basic software for image enhancement can be developed relatively easily, for example with the Java Advanced Imaging library (JAI), and any purpose-built software to quantify scribal hands can (and should) reveal its algorithms and log every detail of what it does.

Regarding understanding and scholarly engagement, it is unreasonable to expect palaeographers to understand the detailed mathematics of image-processing and data-mining, but the techniques are promising and should not be discarded lightly. One possibility is therefore to find ways of presenting the results of this software in ways that are intelligible to palaeographers: if the results can be understood then they are more likely to be accepted. This problem is complex, though, as data-mining often produces results which are not intelligible and cannot necessarily be articulated in an intelligible way.

The Framework for Image Analysis

To test some of these principles, the author of the present paper has developed new modular and extendible software in Java for the analysis of scribal hands. Users can load images of handwriting and then run processes to generate metrics for scribal hands, where each process implements one or more algorithms to extract features from images of handwriting and records both the process and the result. A module is also provided to enhance images before processing, and this can be run as a stand-alone application to help recover text from damaged manuscripts (Stokes, forthcoming). The system can compare metrics generated by different processes and thereby measure distances between samples of writing, and both metrics and distances can be exported for further study. The processes are implemented as plugins so that users can choose different combinations of processes for different situations and can also implement their own algorithms and exchange these to allow others to test their methods and reproduce their results. This allows people to compare different techniques in a common framework, producing libraries of scribal hands and plugins as a common and documented basis for palaeographical study.



```
Hand
file://localhost/Users/pas53/Pictures/MS Images/Other/Antwerp, M.16.2, 48r.JPG
file://localhost/Users/pas53/Pictures/MS Images/Other/Beinecke 516a.jpg
file://localhost/Users/pas53/Pictures/MS Images/Other/Beinecke 516b.jpg
file://localhost/Users/pas53/Pictures/MS Images/Other/Beinecke 516c.jpg
file://localhost/Users/pas53/Pictures/MS Images/Other/2. Connor VIII.tif
file://localhost/Users/pas53/Pictures/MS Images/Oxford/Junius 121, 15v.jpg
```

```
[0.002746743613911512, 0.24149971253626534, 11.907810963841804, 0.0
file://localhost/Users/pas53/Pictures/MS Images/Other/Beinecke 516b.jpg
file://localhost/Users/pas53/Pictures/MS Images/Other/Beinecke 516c.jpg
Autocorrelation (Raw)
Edge directions
Horizontal runs
Autocorrelation
Vertical runs
[2.0687196220426507E-4, 0.023966273633395794, 8.604343123447059, 0.
```

Fig. 1: Framework for Image Analysis

A system like this should alleviate some of the problems discussed above. All analysis using the framework is documented and reproducible, assuming that the images are freely accessible and that the plugins conform to the principles outlined above. The results provide objective evidence, the validity and interpretation of which can then be debated. The metrics and processes may or may not be meaningful to scholars in the humanities, depending on the plugins, but at least they are open to scrutiny by those who can and wish to do so. Indeed, it is an important question how the results of complex algorithms can best be presented to scholars in the humanities, and it may well be that the plugins should allow both 'computer-friendly' and 'human-friendly' output, with the latter including graphical or even interactive displays. Much work is still required here, but the system described is designed for precisely this sort of experimentation and thereby, one hopes, provides a useful environment in which the work can be done.

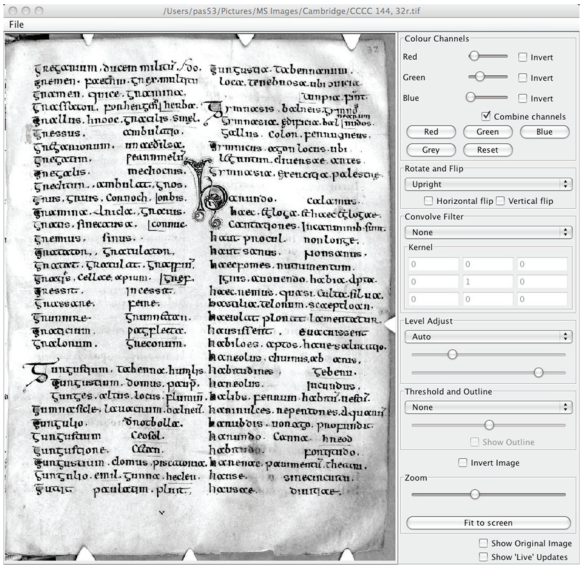


Fig. 2: The Image Enhancement module.
Manuscript page is Cambridge, Corpus Christi College 144, 32r. Reproduced by permission of the Master and Fellows, Corpus Christi College Cambridge

References

Bulacu, M., and Schomaker, L. (2007). Automatic Handwriting Identification on Medieval Documents. *Proc. of 14th Int. Conf. on Image Analysis and Processing* (ICIAP 2007), pp. 279–284. <http://www.ai.rug.nl/~bulacu/iciap2007-bulacu-schomaker.pdf>

Ciula, A. (2005). Digital Palaeography: Using the Digital Representation of Medieval Script to Support Palaeographic Analysis. *Digital Medievalist* 1. <http://www.digitalmedievalist.org/journal/1.1/ciula/>

Costamagna, G., *et al*. (1995 and 1996). Commentare Bischoff. *Scrittura e Civiltà* 19: 325–48 and 20:401–7.

Craig-McFeely, J., and Lock, A. (2006). *Digital Image Archive of Medieval Music: Digital Restoration Workbook*. Oxford: Oxford Select Specialist Catalogue Publications. <http://www.methodsnetwork.ac.uk/redist/pdf/workbook1.pdf>

Davis, T. (2007). The practice of Handwriting Identification. *The Library (7th Series)* 8: 251–76.

Derolez, A. (2003). *The Palaeography of Gothic Manuscript Books*. Cambridge: Cambridge University Press.

Gilissen, L. (1973). *L'expertise des écritures médiévales*. Ghent: Éditions scientifiques E. Story-Scientia.

Muzerelle, D. (2003). *Vocabulaire codicologique*. <http://vocabulaire.irht.cnrs.fr/vocab.htm>

Pierazzo, E. (2008). Editorial Teamwork in a Digital Environment: The Edition of the Correspondence of Giacomo Puccini. *Jahrbuch für Computerphilologie* 10. <http://computerphilologie.tu-darmstadt.de/jg08/pierazzo.html>

Rumble, A. R. (2005). Palaeography, Scribal Identification and the Study of Manuscript Characteristics. *Care and Conservation of Manuscripts: Proceedings of the 8th International Seminar*, edited by G. Fellows-Jensen and P. Springborg. Copenhagen: Museum Tusculanum Press, pp. 217–28.

Scragg, D. G., *et al*. (2005). *ManCASS C11 Database Project*. <http://www.arts.manchester.ac.uk/mancass/C11database/>

Srihari, S. N., *et al*. (2002). Individuality of Handwriting. *Journal of Forensic Science* 47: 1–17.

Stokes, P. A. (2007/8). Palaeography and Image-Processing: Some Solutions and Problems. *Digital Medievalist* 3. <http://www.digitalmedievalist.org/journal/3/stokes/>

Stokes, P. A. (forthcoming). Recovering Anglo-Saxon Erasures: Some Questions, Tools and Techniques. *Palimpsests and the Literary Imagination of Medieval England*, edited by R. Chai-Elsholz *et al*. New York: Palgrave.

Rehbein, M., *et al*., eds. (forthcoming). *Kodikologie und Paläographie im Digitalen Zeitalter—Codicology and Palaeography in the Digital Age*. Norderstedt: Books on Demand.

All URLs last accessed 12 March 2009.

"Terminal Hopscotch": Navigating Networked Space in Talan Memmott's *Lexia to Perplexia*

Lisa Swanstrom
Brandeis University
swanstro@gmail.com

If the reader is patient enough to make it to the final episodes of *Lexia to Perplexia*, digital artist Talan Memmott's beautifully intricate piece of electronic literature, at some point she will encounter a particularly puzzling image. Richly layered with images of computer code, checkerboard backdrops, cryptic prose, and a stick figure drawn in chalk, this snapshot will nevertheless lack the full frenzy of motion of the actual work, a palimpsest-like environment that threatens to spin out of control. The reader will participate in its wild oscillation by moving her mouse around the screen in an effort to find a gateway or portal to the next segment. By this time she will (probably) have learned that this type of active searching is the only way to proceed through the text. For her efforts, she will not be rewarded with a new section, a sense of closure, nor formal dénouement; rather, her participation here will involve playing what Memmott has referred to as "terminal hopscotch," a looping sequence of animations that unfold ad infinitum, or at least until the reader will choose to withdraw.

Created for the trAce Online Writing Community's annual Conference in 2000, *Lexia to Perplexia* is comprised of four sections divided into a series of thickly-layered web pages, each of which leads to further layers before linking to a new page. The text is a mixture of DHTML and Javascript, which, when strung together, forms a fragmented narrative that is visually complemented by empty grids, snippets of code, and cluttered signs of death and mourning. Like many examples of electronic literature or digital poetry, *Lexia to Perplexia* emerges from a variety of artistic traditions: it is a visual poem, a linguistic experiment, and a piece of executable code all in one. It is a technological collage performed on a computer screen, filled with references to various media forms while settling exclusively on none.

Yet there are many things that *Lexia to Perplexia* is not. It is not a linear narrative, yet through a sustained interaction with the piece, an abstract sort of "story" emerges. It does not contain a set series of causal sequences, yet any path through this image-laden text is causally dependent upon the decisions the user makes in response to its interface. It is not a story in which a single "main character" functions as a center of narration, yet the piece is suggestive of a subjective amalgam, a stitched-together entity that has joined itself to a network. Nor does the piece fit entirely within the tradition of the avant-garde in the visual arts—even as its operative and interactive features express an unusual sort of seeing, a "networked perspective," that resonates with Marcel Duchamp's attempts to deter the "retinal shudder" that results from traditional manners of representation (qtd. in Ades 70).

In her detailed analysis of the work in *Writing Machines*, N. Katherine Hayles discusses the manner in which Memmott creates an anonymous protagonist resigned to a divided and encoded existence that the computer interface and distributed network have made inevitable. Her convincing argument is that in *Lexia to Perplexia* "human subjectivity is depicted as intimately entwined with computer technologies," and that such an entwinement is achieved through Memmott's use of "idiosyncratic language, a revisioning of classical myths, and a set of coded images that invite the reader to understand herself not as a preexisting self with secure boundaries but as a permeable membrane through which information flows" (50). Hayles' analysis not only addresses the unique character of Memmott's project, it does so in such a way as to shed light upon digital art as a whole, providing ample evidence to support the one of the theses that underpins *Writing Machines*: the increasing need for media-specific analysis.

Yet far from exhausting what can be gleaned from *Lexia to Perplexia*, Hayles' clear and thorough exegesis creates a new mode of approach. If, for example, it is crucial to apply a media-specific analysis to works such as *Lexia to Perplexia*, then the task remains to consider more fully one of its most important medial features: the organization of spatial elements on the level of the interface. In this paper I depart from earlier modes of interpretation that focus primarily on linguistic elements (see, for example, work by Hayles, Dreher, and Raley) by considering the way Memmott complicates traditional notions of identity in *Lexia to Perplexia* through a unique arrangement of spatial elements that form a networked reading environment and suggest a distributed subjectivity. Such an arrangement creates an anxious landscape that offers no place for repose, one that, at least initially, makes a refugee of the reader as she "hopscotches" through the text. Additionally, I propose that *Lexia to Perplexia*'s distinct spatial organization results in a form of storytelling that both challenges and conforms to traditional narrative structure—and that these spatial elements, dependent as they are upon the medium of the computer and the technology of the distributed network, engender new ways of navigating textual space that demand new literacies to fully experience them.

More about *gentleman* in Dickens

Tomoji Tabata
The University of Osaka
tabata@lang.osaka-u.ac.jp

Introduction

The aim of this paper is to present a corpus-stylistic study of the collocation of *gentleman* in Dickens. The word *gentleman* is among the most frequent 'content' words in Dickens. In fact, as Fig. 1 shows, *gentleman* appears more frequently in Dickens than in any other author examined, and thus is a *key* word in Dickens in the sense that it 'appear[s] in a text or a part of a text with a frequency greater than chance occurrence alone would suggest' (Henry and Rooseberry, 2001: 110).

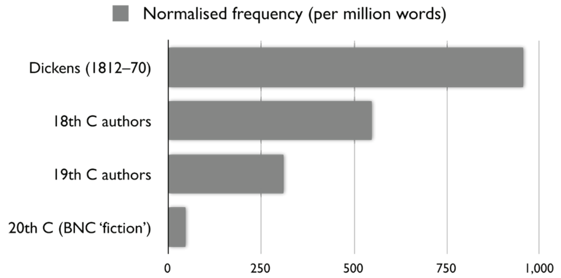


Fig. 1 Normalised frequency of gentleman *per million words*

Building on my pilot study of the collocation of *gentleman* in Dickens and Smollett (Tabata, 2008), this study expands the scope of analysis by comparing Dickens texts with a larger reference corpus covering major eighteenth- and nineteenth-century authors as well as by combining quantitative techniques with qualitative interpretation of statistical findings. The corpus, upon which the present study is based, is made up of three components: 1) 24 Dickens text sets (4,835,158 words), 2) a set of 23 eighteenth-century texts (4,163,353 words: Defoe, Fielding, Goldsmith, Richardson, Smollett, Sterne, and Swift), and 3) a set of 31 nineteenth-century texts (5,118,346 words: Austen, the Brontë sisters, Collins, George Eliot, Gaskell, Thackeray, and Trollope). The total of running words amount to 14,116,857. One might find that female authors outnumber male authors in the nineteenth-century set. However, the female set is not so overpopulated as to imbalance the population of the subcorpus. The total of the tokens by male authors accounts for as high as 45 % of the running words due to comparatively thick volumes produced by male authors.

A few methodological issues

To investigate Dickens' stylistic features associated with the use of *gentleman*, it will be appropriate to analyze the word in collocation, rather than in isolation since semantics of a word is extended to the surrounding words, or co-text (Firth, 1957; Sinclair, 1991; Stubbs, 2001). The first issue to be discussed is a collocational span. Although there is not a total agreement between linguists regarding an optimal range of collocational span, a generally accepted practice is to examine collocation in a span of four words to the each side of the node (Jones and Sinclair, 1974). This is based on the finding that it will become increasingly diffcult to find meaningful collocational pattens beyond a span of four words (Sinclair *et al*., 2004), as shown in Fig. 2.

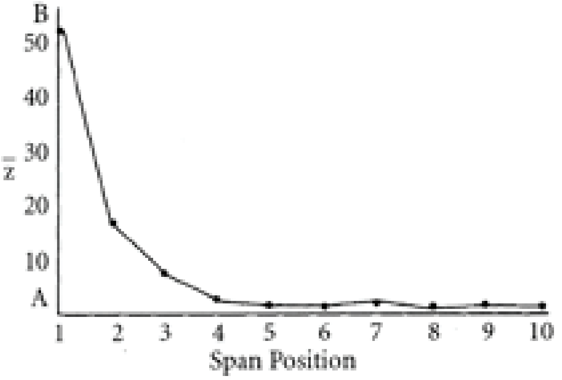


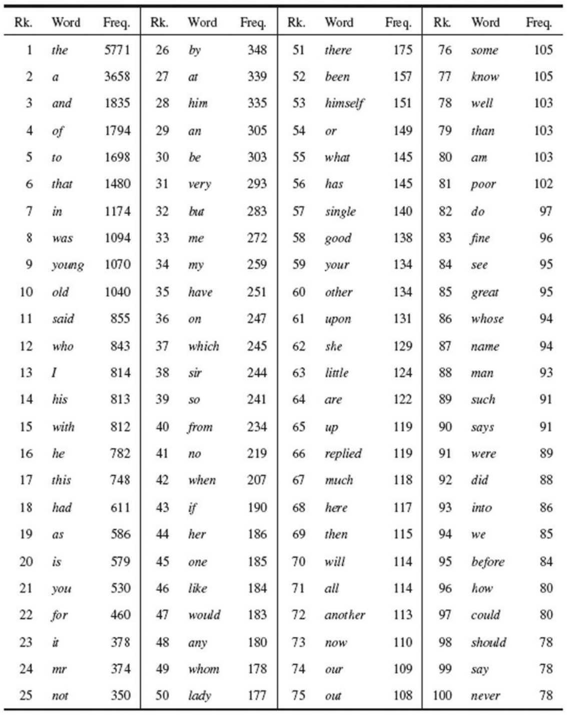
*Fig. 2 Graph showing average node predictions over span positions 1–10**
** From Sinclair (1969) as reprinted in Sinclair et al. (2004)*

Next comes the issue of how to measure collocational strength between words. If we use raw frequency (or normalised frequency) counts, the predominance of function words such as *the, a, and, of*, etc. (as evident in Table 1) would overcrowd subtler, more meaningful patterns. My proposed solution is to use a statistical measure. Among a number of techniques for filtering out unimportant neighbouring words, a Mutual Information (MI) score measures collocational strength, by a logarithmic compression of the frequency of collocates. It therefore is likely to spotlight semantic relationships rather than syntactic relationships between the node and its collocates. Mutual information score for the collocation of the word *x* and the word *y* ($I_{(x,y)}$) is obtained from the formula (1).

$$I_{(x,y)} = \log_2 \frac{f_{(x,y)} \quad N}{f_{(x)} \quad f_{(y)}}$$

The third issue is how we set a threshold for variables. Church and Hanks (1990: 24) state that the association ratio becomes unstable when the counts are very small (for example, when $f_{x.y} \leq 5$). My tentative proposal is to base analysis upon collocates occurring 10 times or more, and with MI scores higher than 3.0, drawing on Church and Hanks' account that "pairs with $I_{(x,y)} > 3$ tend to be interesting" (24). Table 2 list 100 strongest collocates with $f_{(x,y)} \geq 10$. A close look at Table 2, however, reveals that there are a few proper nouns that occur only in a particular text (Brass's, Greystock, and Oliver). As a results of eliminating twelve such proper nouns from the candidates for variables, 378 collocates of *gentleman* were found qulified as variables. The initial number of texts in the corpus was 78, but one text, Smollett's *The History and Adventures of an Atom* (1769) did not meet the requirements and, therefore, was eliminated from the data set.

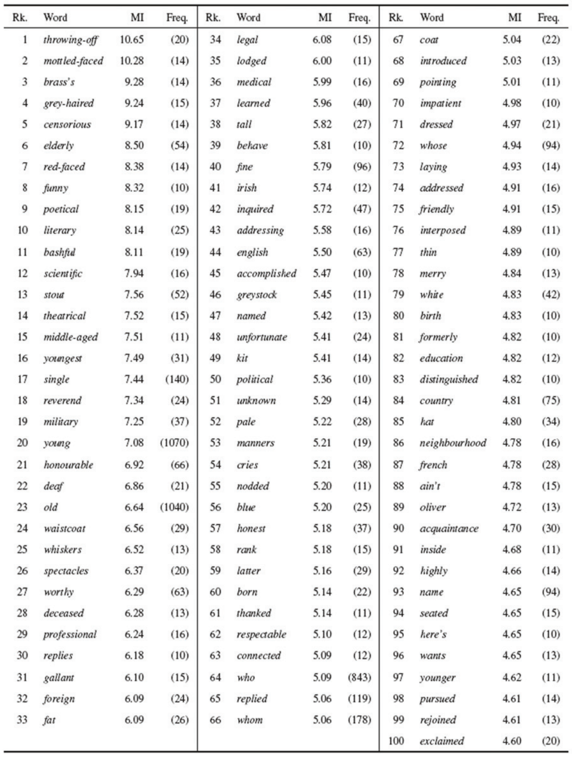
The sheer volume of data to be analysed is daunting: a collocational strength matrix for 378 collocates across 77 texts (77 rows by 378 columns). Since it would be extremely diffcult for a human eye to detect meaningful patterns from the data of such dimensions, correspondence analysis is employed to visualize complex interrelationships among *gentleman*'s collocates, interrelationships among texts, and the association patterns between the *gentleman*'s collocates and texts in multidimensional spaces.



Rk.	Word	Freq.	Rk.	Word	Freq.	Rk.	Word	Freq.	Rk.	Word	Freq.
1	*the*	5771	26	*by*	348	51	*there*	175	76	*some*	105
2	*a*	3658	27	*at*	339	52	*been*	157	77	*know*	105
3	*and*	1835	28	*him*	335	53	*himself*	151	78	*well*	103
4	*of*	1794	29	*an*	305	54	*or*	149	79	*than*	103
5	*to*	1698	30	*be*	303	55	*what*	145	80	*am*	103
6	*that*	1480	31	*very*	293	56	*has*	145	81	*poor*	102
7	*in*	1174	32	*but*	283	57	*single*	140	82	*do*	97
8	*was*	1094	33	*me*	272	58	*good*	138	83	*fine*	96
9	*young*	1070	34	*my*	259	59	*your*	134	84	*see*	95
10	*old*	1040	35	*have*	251	60	*other*	134	85	*great*	95
11	*said*	855	36	*on*	247	61	*upon*	131	86	*whose*	94
12	*who*	843	37	*which*	245	62	*she*	129	87	*name*	94
13	*I*	814	38	*sir*	244	63	*little*	124	88	*man*	93
14	*his*	813	39	*so*	241	64	*are*	122	89	*such*	91
15	*with*	812	40	*from*	234	65	*up*	119	90	*says*	91
16	*he*	782	41	*no*	219	66	*replied*	119	91	*were*	89
17	*this*	748	42	*when*	207	67	*much*	118	92	*did*	88
18	*had*	611	43	*if*	190	68	*here*	117	93	*into*	86
19	*as*	586	44	*her*	186	69	*then*	115	94	*we*	85
20	*is*	579	45	*one*	185	70	*will*	114	95	*before*	84
21	*you*	530	46	*like*	184	71	*all*	114	96	*how*	80
22	*for*	460	47	*would*	183	72	*another*	113	97	*could*	80
23	*it*	378	48	*any*	180	73	*now*	110	98	*should*	78
24	*mr*	374	49	*whom*	178	74	*our*	109	99	*say*	78
25	*not*	350	50	*lady*	177	75	*out*	108	100	*never*	78

Table 1 100 most common collocates of gentleman

Various multivariate analyses of texts have been successful in elucidating linguistic variation over time, variation across registers, variation across oceans, to say nothing of linguistic differences between authors (Brainerd, 1980; Burrows, 1987 & 1996; Biber and Finegan, 1992; Craig, 1999a, b, & c; Hoover, 2003a, b, & c; Rudman, 2005). My earlier attempts used correspondence analysis to accommodate low frequency variables (words) in profiling authorial/chronological/cross-register variations in Dickens and Smollett (Tabata, 2005; 2007a/b; 2008). Given the fact that most collocates of content words tend to be low in frequency, my methodology based on correspondence analysis would usefully be applied to a macroscopic analysis of collocation of *gentleman*.



Rk.	Word	MI	Freq.	Rk.	Word	MI	Freq.	Rk.	Word	MI	Freq.
1	*throwing-off*	10.65	(20)	34	*legal*	6.08	(15)	67	*coat*	5.04	(22)
2	*mottled-faced*	10.28	(14)	35	*lodged*	6.00	(11)	68	*introduced*	5.03	(13)
3	*brass's*	9.28	(14)	36	*medical*	5.99	(16)	69	*pointing*	5.01	(11)
4	*grey-haired*	9.24	(15)	37	*learned*	5.96	(40)	70	*impatient*	4.98	(10)
5	*censorious*	9.17	(14)	38	*tall*	5.82	(27)	71	*dressed*	4.97	(21)
6	*elderly*	8.50	(54)	39	*behave*	5.81	(10)	72	*whose*	4.94	(94)
7	*red-faced*	8.38	(14)	40	*fine*	5.79	(96)	73	*laying*	4.93	(14)
8	*funny*	8.32	(10)	41	*irish*	5.74	(12)	74	*addressed*	4.91	(16)
9	*poetical*	8.15	(19)	42	*inquired*	5.72	(47)	75	*friendly*	4.91	(15)
10	*literary*	8.14	(25)	43	*addressing*	5.58	(16)	76	*interposed*	4.89	(11)
11	*bashful*	8.11	(19)	44	*english*	5.50	(63)	77	*thin*	4.89	(10)
12	*scientific*	7.94	(16)	45	*accomplished*	5.47	(10)	78	*merry*	4.84	(13)
13	*stout*	7.56	(52)	46	*greystock*	5.45	(11)	79	*white*	4.83	(42)
14	*theatrical*	7.52	(15)	47	*named*	5.42	(13)	80	*birth*	4.83	(10)
15	*middle-aged*	7.51	(11)	48	*unfortunate*	5.41	(24)	81	*formerly*	4.82	(10)
16	*youngest*	7.49	(31)	49	*kit*	5.41	(14)	82	*education*	4.82	(12)
17	*single*	7.44	(140)	50	*political*	5.36	(10)	83	*distinguished*	4.82	(10)
18	*reverend*	7.34	(24)	51	*unknown*	5.29	(14)	84	*country*	4.81	(75)
19	*military*	7.25	(37)	52	*pale*	5.22	(28)	85	*hat*	4.80	(34)
20	*young*	7.08	(1070)	53	*manners*	5.21	(19)	86	*neighbourhood*	4.78	(16)
21	*honourable*	6.92	(66)	54	*cries*	5.21	(38)	87	*french*	4.78	(28)
22	*deaf*	6.86	(21)	55	*nodded*	5.20	(11)	88	*ain't*	4.78	(15)
23	*old*	6.64	(1040)	56	*blue*	5.20	(25)	89	*oliver*	4.72	(13)
24	*waistcoat*	6.56	(29)	57	*honest*	5.18	(37)	90	*acquaintance*	4.70	(30)
25	*whiskers*	6.52	(13)	58	*rank*	5.18	(15)	91	*inside*	4.68	(11)
26	*spectacles*	6.37	(20)	59	*latter*	5.16	(29)	92	*highly*	4.66	(14)
27	*worthy*	6.29	(63)	60	*born*	5.14	(22)	93	*name*	4.65	(94)
28	*deceased*	6.28	(13)	61	*thanked*	5.14	(11)	94	*seated*	4.65	(15)
29	*professional*	6.24	(16)	62	*respectable*	5.10	(12)	95	*here's*	4.65	(10)
30	*replies*	6.18	(10)	63	*connected*	5.09	(12)	96	*wants*	4.65	(13)
31	*gallant*	6.10	(15)	64	*who*	5.09	(843)	97	*younger*	4.62	(11)
32	*foreign*	6.09	(24)	65	*replied*	5.06	(119)	98	*pursued*	4.61	(14)
33	*fat*	6.09	(26)	66	*whom*	5.06	(178)	99	*rejoined*	4.61	(13)
								100	*exclaimed*	4.60	(20)

Table 2 100 strongest collocates of gentleman *based on MI-score*

Results

Fig. 3 visualises interrelationships among 77 texts. Data points (texts) closer to each other in the diagram tend to have similar collocates in common. The greater the distance between texts, the less they have in common. Fig. 4 indicates interrelationships among 378 collocates. The proportion accounted for by Dimensions 1 and 2 is only 4.21 % and 3.03 %, respectively, of the total variance in the data, indicating the relationships among the matrix of 77 rows by 378 columns are extremely complex.

Fig. 3 provides an interesting overview of similarity or contrast between texts: the horizontal axis, the strongest variance in the data set, differentiates Dickens versus the eighteenth-and nineteenth-century authors. One seeming anomaly as far as Dickens texts are concerned is the position of 1851_CHE, *A Child's History of England* (1851), which finds itself between the eighteenth and the nineteenth century text groups. This history book written for children is considerably different in style from other Dickensian works. Therefore it is not unexpected for this piece to be found least Dickensian as indicated by the vertical axis, a phenomenon in keeping with previous multivariate studies based on other linguistic variables, such as *–ly* adverbs (Tabata, 2005: 231) and part-of-speech distribution (Tabata, 2002: 173). In addition, early Dickens texts are found in the lower half of the Dickens cluster, again in consistent with my other works (Tabata, 2008; 2009a; 2009b).

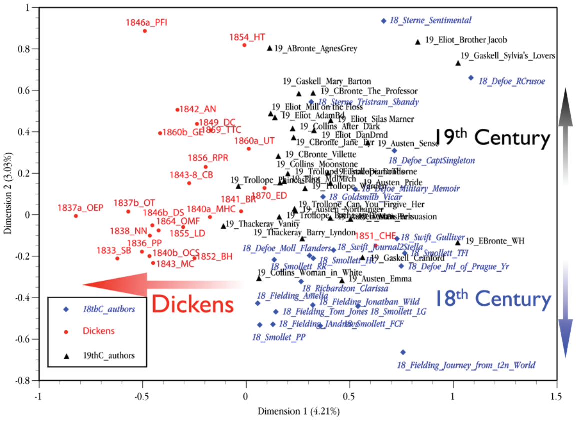


Fig. 3 Correspondence Analysis of the collocates of gentleman (378 collocates across 77 texts)

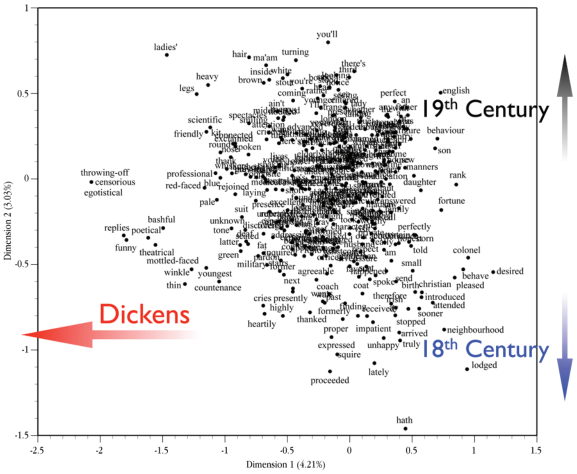
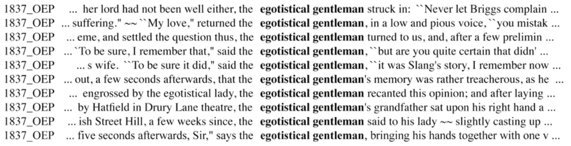


Fig. 4 A galaxy of gentleman's collocates: Word-map of 378 collocates

The vertical axis, furthermore, shows the eighteenth-century texts to wards the bottom and the nineteenth century texts towards the upper half of the chart although the two sets are not in two distinct clusters. Fig. 4 corresponds to Fig. 3 and thus tells us what words tend to co-occur with *gentleman* in Dickens, in the eighteenth- and nineteenth-century texts.

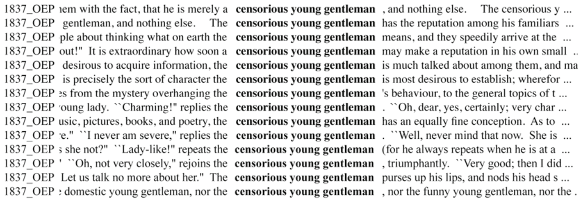
Concordance

A close examination of Fig. 4 leads to an awareness that *gentleman* in Dickens are characterised by (uncommon) adjectives while, in the eighteenth-century texts, verbs (of past tense) are prominent collocates. The nineteenth-century texts do not display a particular pattern apart from words related to family or position. I would rather interpret this result as suggesting the nineteenth-century texts are negatively characterised both against the Dickens set and the eighteenth-century set.



```
1837_OEP    ...  her lord had not been well either, the  egotistical gentleman struck in:  ``Never let Briggs complain ...
1837_OEP  ... suffering." ~~ ``My love," returned the  egotistical gentleman, in a low and pious voice, ``you mistak ...
1837_OEP   ... eme, and settled the question thus, the  egotistical gentleman turned to us, and, after a few prelimin ...
1837_OEP  ... `To be sure, I remember that," said the  egotistical gentleman, ``but are you quite certain that didn' ...
1837_OEP       ... s wife.  ``To be sure it did," said the  egotistical gentleman, ``it was Slang's story, I remember now ...
1837_OEP  ... out, a few seconds afterwards, that the  egotistical gentleman's memory was rather treacherous, as he  ...
1837_OEP    ...  engrossed by the egotistical lady, the  egotistical gentleman recanted this opinion; and after laying ...
1837_OEP  ...  by Hatfield in Drury Lane theatre, the  egotistical gentleman's grandfather sat upon his right hand a ...
1837_OEP   ... ish Street Hill, a few weeks since, the  egotistical gentleman said to his lady ~~ slightly casting up ...
1837_OEP   ... five seconds afterwards, Sir," says the  egotistical gentleman, bringing his hands together with one v ...
```

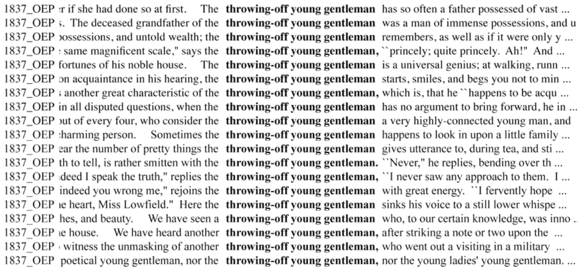
Fig. 5 Concordance: egotistical



```
1837_OEP em with the fact, that he is merely a   censorious young gentleman  , and nothing else.    The censorious y ...
1837_OEP  gentleman, and nothing else.     The   censorious young gentleman  has the reputation among his familiars  ...
1837_OEP ple about thinking what on earth the   censorious young gentleman  means, and they speedily arrive at the  ...
1837_OEP out!"  It is extraordinary how soon a   censorious young gentleman  may make a reputation in his own small  ..
1837_OEP  desirous to acquire information, the   censorious young gentleman  is much talked about among them, and ma
1837_OEP  is precisely the sort of character the   censorious young gentleman  is most desirous to establish; wherefor ...
1837_OEP es from the mystery overhanging the   censorious young gentleman  's behaviour, to the general topics of t ...
1837_OEP oung lady.  ``Charming!" replies the   censorious young gentleman  .  ``Oh, dear, yes, certainly; very char ...
1837_OEP usic, pictures, books, and poetry, the   censorious young gentleman  has an equally fine conception.  As to  ...
1837_OEP e."  ``I never am severe," replies the   censorious young gentleman  .  ``Well, never mind that now.  She is  ...
1837_OEP s she not?"  ``Lady-like!" repeats the  censorious young gentleman  (for he always repeats when he is at a  ...
1837_OEP '  ``Oh, not very closely," rejoins the  censorious young gentleman  , triumphantly.  ``Very good; then I did ...
1837_OEP  Let us talk no more about her."  The  censorious young gentleman  purses up his lips, and nods his head s ...
1837_OEP  domestic young gentleman, nor the  censorious young gentleman  , nor the funny young gentleman, nor the .
```

Fig. 6 Concordance: censorious

A close study of *gentleman* in collocation with such Dickensian adjectives, *egotistical, censorious, throwing-off* makes us realize Dicken's ironical use of *gentleman*. Moreover, these concordance lines make us realize that in Dickens more than one modifiers are likely to precede *gentleman*. In fact, Fig. 8 illustrates Dickens's tendency to use multiple adjectives. The increase in the proportion which Dickens instances occupy in 'the multi-adjective gentleman' from '*a* | *an* ADJ ADJ gentleman', as well as *gentleman* in total indicates that Dickens uses '*the* ADJ ADJ *gentleman*' formula as character appellations, instead of simply referring to a character as *a gentleman*.



```
1837_OEP r if she had done so at first.     The  throwing-off young gentleman  has so often a father possessed of vast ...
1837_OEP s.  The deceased grandfather of the  throwing-off young gentleman  was a man of immense possessions, and u
1837_OEP possessions, and untold wealth; the  throwing-off young gentleman  remembers, as well as if it were only y ...
1837_OEP  same magnificent scale," says the  throwing-off young gentleman, ``princely; quite princely.  Ah!"  And ...
1837_OEP fortunes of his noble house.     The  throwing-off young gentleman  is a universal genius; at walking, runn ...
1837_OEP on acquaintance in his hearing, the  throwing-off young gentleman  starts, smiles, and begs you not to min ...
1837_OEP  another great characteristic of the  throwing-off young gentleman, which is, that he ``happens to be acqu ...
1837_OEP in all disputed questions, when the  throwing-off young gentleman  has no argument to bring forward, he in ...
1837_OEP out of every four, who consider the  throwing-off young gentleman  a very highly-connected young man, and
1837_OEP harming person.     Sometimes the  throwing-off young gentleman  happens to look in upon a little family ...
1837_OEP ear the number of pretty things the  throwing-off young gentleman  gives utterance to, during tea, and sti ...
1837_OEP th to tell, is rather smitten with the  throwing-off young gentleman. ``Never," he replies, bending over th ...
1837_OEP deed I speak the truth," replies the  throwing-off young gentleman, ``I never saw any approach to them.  I ...
1837_OEP indeed you wrong me," rejoins the  throwing-off young gentleman  with great energy.  ``I fervently hope  ...
1837_OEP e heart, Miss Lowfield."  Here the  throwing-off young gentleman  sinks his voice to a still lower whispe ...
1837_OEP hes, and beauty.     We have seen a  throwing-off young gentleman  who, to our certain knowledge, was inno ..
1837_OEP e house.     We have heard another  throwing-off young gentleman, after striking a note or two upon the  ...
1837_OEP  witness the unmasking of another  throwing-off young gentleman, who went out a visiting in a military  ...
1837_OEP  poetical young gentleman, nor the  throwing-off young gentleman, nor the young ladies' young gentleman. ...
```

Fig. 7 Concordance: throwing-off

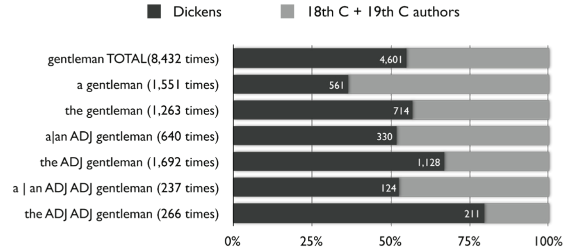
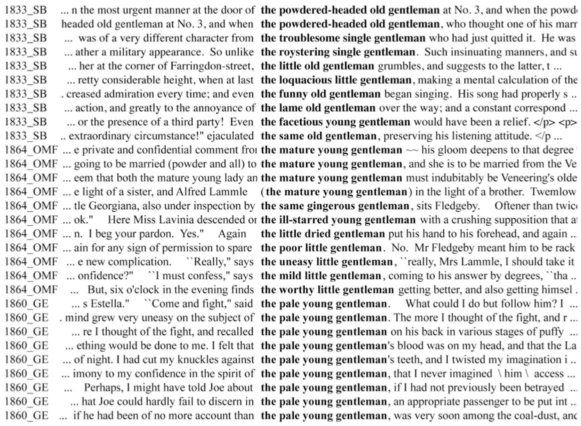
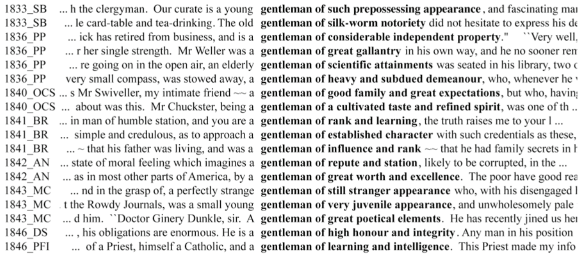


Fig. 8 Proportion of instances accounted for by Dickens



```
1833_SB    ... n the most urgent manner at the door of  the powdered-headed old gentleman at No. 3, and when the powd
1833_SB    headed old gentleman at No. 3, and when  the powdered-headed old gentleman, who thought one of his marr
1833_SB     ...  was of a very different character from  the troublesome single gentleman who had just quitted it.  He was
1833_SB     ... ather a military appearance.  So unlike  the roystering single gentleman.  Such insinuating manners, and su
1833_SB      ... her at the corner of Farringdon-street,  the little old gentleman grumbles, and suggests to the latter, t ...
1833_SB      ... retty considerable height, when at last  the loquacious little gentleman, making a mental calculation of the
1833_SB   . creased admiration every time; and even  the funny old gentleman began singing.  His song had properly s ..
1833_SB     ... action, and greatly to the annoyance of  the lame old gentleman over the way; and a constant correspond ...
1833_SB      ... or the presence of a third party!  Even  the facetious young gentleman would have been a relief. </p> <p>
1833_SB   .. extraordinary circumstance!" ejaculated  the same old gentleman, preserving his listening attitude. </p ...
1864_OMF ... e private and confidential comment fro  the mature young gentleman ~~ his gloom deepens to that degree
1864_OMF ... going to be married (powder and all) to  the mature young gentleman, and she is to be married from the Ve
1864_OMF ... eem that both the mature young lady an  the mature young gentleman must indubitably be Veneering's olde
1864_OMF ... e light of a sister, and Alfred Lammle    (the mature young gentleman) in the light of a brother.  Twemlow
1864_OMF ... tle Georgiana, also under inspection by  the same gingerous gentleman, sits Fledgeby.    Oftener than twic
1864_OMF ... ok."     Here Miss Lavinia descended or  the ill-starred young gentleman with a crushing supposition that a
1864_OMF ... n.  I beg your pardon.  Yes."     Again    the little dried gentleman put his hand to his forehead, and again ...
1864_OMF ... ain for any sign of permission to spare   the poor little gentleman.  No.  Mr Fledgeby meant him to be rack
1864_OMF   ... e new complication.     ``Really," says  the uneasy little gentleman, ``really, Mrs Lammle, I should take it
1864_OMF   ... onfidence?"     ``I must confess," says  the mild little gentleman, coming to his answer by degrees, ``tha ..
1864_OMF    ...   But, six o'clock in the evening finds  the worthy little gentleman getting better, and also getting himsel .
1860_GE       ... s Estella."     ``Come and fight," said  the pale young gentleman.     What could I do but follow him? I  ...
1860_GE    . mind grew very uneasy on the subject of  the pale young gentleman. The more I thought of the fight, and r ...
1860_GE       ... re I thought of the fight, and recalled  the pale young gentleman on his back in various stages of puffy  ...
1860_GE     ... ething would be done to me. I felt that  the pale young gentleman's blood was on my head, and that the La
1860_GE    ... of night. I had cut my knuckles against  the pale young gentleman's teeth, and I twisted my imagination i ..
1860_GE    ... imony to my confidence in the spirit of  the pale young gentleman, that I never imagined  \ him \  access ...
1860_GE      ...    Perhaps, I might have told Joe about  the pale young gentleman, if I had not previously been betrayed  ...
1860_GE       ... hat Joe could hardly fail to discern in  the pale young gentleman, an appropriate passenger to be put int ...
1860_GE    ...  if he had been of no more account than  the pale young gentleman, was very soon among the coal-dust, anc
```

Fig. 9 Concordance: the X X gentleman



```
1833_SB    ... h the clergyman.  Our curate is a young  gentleman of such prepossessing appearance, and fascinating mai
1833_SB     ... le card-table and tea-drinking. The old  gentleman of silk-worm notoriety did not hesitate to express his de
1836_PP      ... ick has retired from business, and is a  gentleman of considerable independent property."     ``Very well,
1836_PP     ... r her single strength.  Mr Weller was a  gentleman of great gallantry in his own way, and he no sooner rem
1836_PP      ... re going on in the open air, an elderly  gentleman of scientific attainments was seated in his library, two c
1836_PP     very small compass, was stowed away, a  gentleman of heavy and subdued demeanour, who, whenever he
1840_OCS ... s Mr Swiveller, my intimate friend ~~ a  gentleman of good family and great expectations, but who, havin
1840_OCS ...  about was this.  Mr Chuckster, being a  gentleman of a cultivated taste and refined spirit, was one of th ..
1841_BR   ... in man of humble station, and you are a  gentleman of rank and learning, the truth raises me to your l ...
1841_BR   ...  simple and credulous, as to approach a  gentleman of established character with such credentials as these,
1841_BR     ... ~ that his father was living, and was a  gentleman of influence and rank ~~ that he had family secrets in h
1842_AN   ... state of moral feeling which imagines a  gentleman of repute and station, likely to be corrupted, in the ...
1842_AN    ... as in most other parts of America, by a  gentleman of great worth and excellence.  The poor have good rea
1843_MC      ... nd in the grasp of, a perfectly strange  gentleman of still stranger appearance who, with his disengaged l
1843_MC    . t the Rowdy Journals, was a small young  gentleman of very juvenile appearance, and unwholesomely pale
1843_MC    ... d him.  ``Doctor Ginery Dunkle, sir.  A  gentleman of great poetical elements.  He has recently jined us her
1846_DS      ... , his obligations are enormous. He is a  gentleman of high honour and integrity. Any man in his position
1846_PFI      ...  of a Priest, himself a Catholic, and a  gentleman of learning and intelligence.  This Priest made my info
```

Fig. 10 Concordance: gentleman of X X

Conclusion

What has emerged from this survey can be summarised as follows:

1. The most powerful solution obtained from multivariate analysis can be interpreted as demonstrating that Dickens has distinctive style in collocation of the word *gentleman*.

2. Dickens is more likely to use adjectives in collocation with gentleman than the control set. Adjectives with higher MI scores often strike oxymoronic humour when collocating with *gentleman* in Dickens texts.

3. In Dickens, modifier-collocates (adjectives) tend to occur in succession (juxtaposition or concatenation) typically in early works. They are often employed as character appellation (variations of character-name)

with a negative semantic prosody.

References

Adolphs, S. and R. A. Carter (2003) ‘Corpus stylistics: point of view and semantic prosodies in *To The Lighthouse*’, *Poetica*, 58: 7–20.

Church K. W. and Hunks P. (1990) “Word Association Norms, Mutual Information and Lexicography,” *Computational Linguistics*, 16/1: 22-29.

Firth, J. R. (1935). The Technique of Semantics. *Transactions of the Philological Society*, 36– 72 (Reprinted in Firth (1957) *Papers in Linguistics*. London: Oxford University Press, 7–33).

Firth, J. R. (1957) ‘Modes of Meaning’, in *Papers in Linguistics 1934–51*. London: OUP. 191-215.

Greenbaum, S. (1970) *Verb-Intensifier Collocations in English: An experimental approach*. The Hague: Mouton.

Henry. A. and R. L. Rooseberry (2001). Using a small corpus to obtain data for teaching a genrein M. Ghadessy, A. Henry and R. L. Roseberry (eds.) *Small Corpus and ELT*. Amsterdam/Philadelphia, Pa.: John Benjamins. 93–133.

Hori, M. (2004) *Investigating Dickens' Style: A Collocational Analysis*. New York: Palgrave Macmillan.

Hunston, S. and G. Framcis (1999) *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Jones, S. and J. Sinclair (1974) “English Lecical Collocation”, *Cahiers de Lexicologie*, 24: 15–61.

Kjellmer, G. (1994) *A Dictionary of English Collocations: Based on the Brown Corpus (3 vols.)*. Oxford: Clarendon Press.

Louw, W. (1993) ‘Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies’, reprinted in G. Sampson and D. McCarthy (eds.) (2004) *Corpus Linguistics: Readings in a Widening Discipline*. London and New York: Continuum. 229–241.

Partington, A. (2003) *The Linguistics of Political Argument: The Spin-doctor and the Wolf-pack at the White House*. London/New York: Routledge.

Partington, A. (2006) *The Linguistics of Laughter: A Corpus-Assisted Study of Laughter-Talk*. Lon-don/New York: Routledge.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J. M., Jones, S. and R. Daley (2004) *English Collocation Studies*: *The OSTI Report*. Continuum.

Stubbs, M. (1995) ‘Corpus evidence for norms of lexical collocation’, in G. Cook and B. Seidlhofer (eds.) *Principle and Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson*. Oxford: OUP. 243–256.

Stubbs, M. (2001) *Words and Phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Tabata, T. (1995) ‘Narrative Style and the Frequencies of Very Common Words: A Corpus-based Approach to Dickens’s First-person and Third-person Narratives’, *English Corpus Studies*, 2: 91–109.

Tabata, T. (2002) ‘Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution’, in T. Saito, J. Nakamura and S. Yamasaki (eds.) *English Corpus Linguistics in Japan*. Amsterdam: Rodopi. 165–182.

Tabata, T. (2004) ‘Differentiation of Idiolects in Fictional Discourse: A Stylo-Statistical Approach to Dickens’s Artistry’, in R. Hiltunen and S. Watanabe (eds.) *Approaches to Style and Discourse in English*. Osaka: Osaka University Press. 79–106.

Tabata, T. (2005) ‘Profiling stylistic variations in Dickens and Smollett through correspondence analysis of low frequency words’, *ACH/ALLC 2005 Conference Abstracts*, Humanities Computing and Media Centre, University of Victoria, Canada, 229–232.

Tabata, T. (2008) *Gentleman* in Dickens: A multivariate stylometric approach to its collocation, *Digital Humanities 2008, Book of Abstracts, The 20th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, University of Oulu, Finland, June 24–June 29, 2008*, 199–202.

Tabata, T. (2009a) ‘“*Wickedly, Falsely, Traitorously*, and otherwise *Evil-adverbiously*, Revealing” the Author’s Style: Correspondence Analysis of *–ly* adverbs in Dickens and Smollett’, *Stylistic Studies of Literature*: *In Honour of Professor Hiroyuki Ito*. Bern: Peter Lang. 113–134.

Tabata, T. (2009b) '"The *Cunningest, Rummest, Superlativest* Old Fox": A multivariate approach to superlatives in Dickens and Smollett', *English Philology and Corpus Studies*: *A Festschrift in Honor of Professor Mitsunori Imai*. Tokyo: Eihosha. 225–240.

Watanabe, S. (2009) '*Gentleman* in *Oliver Twist*: A Linguistic Approach to Literature', *English Philology and Corpus Studies*: *A Festschrift in Honor of Professor Mitsunori Imai*. Tokyo: Eihosha. 273–286.

Digital Curiosities: Resource Creation via Amateur Digitisation

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

Introduction

Digitisation, "the conversion of an analog signal or code into a digital signal or code" (Lee 2002, 3) is now commonplace in most memory institutions, as digital representations of cultural and historical documents, artefacts, and images are created and delivered to users, generally online. The exponential growth in digitisation projects towards the close of the 20th Century, along with the establishment of guides to good practice and technical guidelines, has meant that "Countless millions of pounds, dollars, francs and marks [have been] ploughed into digital projects that have involved the conversion of library, museum and archive collections" (Lee 2002, 160). Much of the early academic debates regarding the purpose, merit, and scope of digitisation are now resolved as institutions create high quality resources for the general user and academic researcher alike (Hughes 2002, Deegan and Tanner 2002). As a result

> Digitisation is not a per-se research issue but is part of a wider context related to the information society and the effective use of the digital content by cultural institutions (Minerva 2003, xxiii).

However, an area seldom considered in academic literature is the creation of digital resources by amateurs. Although hitherto ignored by information professionals, recent developments in Web 2.0 technologies means that museums, libraries and archives are now re-considering their relationship with users and the general public, both in the use of digital collections and how users can contribute to an increasingly rich digital resource environment. This paper assesses the scope of online resources created outside institutional boundaries by keen individuals who wish to participate in digitising our cultural heritage, providing an overview and conceptualising the potential contribution that can be made by amateur digitisation.

Context

The rise of online "museums" created by amateur enthusiasts, generally containing digital images of holdings and artefacts, is a seldom considered but growing phenomenon. Amateur online collections have appropriated a variety of technologies, from static HTML, to the hosting opportunities afforded by online, new media, social networking sites such as www.flickr.com. In addition, with memory institutions appropriating Web 2.0 technologies themselves – such as tagging, and encouraging user feedback and involvement – amateur enthusiasts are now being encouraged to contribute to the online presence of established institutions. Online "museum" material resulting from amateur digitisation projects can provide a rich source of primary resources for both scholars and the general public, and although this has been all but ignored until recently by the Library, Archive, Cultural Heritage, and Arts and Humanities communities, its democratising nature is worthy of further consideration. "On and on it goes – acres of the cyberworld full of ephemera. What else is out there?" (Gorman, 2003, p. 11).

It is acknowledged that "cyberspace is littered with the productions of ignorant, semi-literate, and/or crazed individuals", (Gorman 2003, p. 14) and in many cases, these online collections function as 21st Century cabinets of curiosities. They can be viewed as amusing, eccentric, or even worrying obsessions with a particular type of ephemera which the rest of the world has chosen to leave undocumented, providing a "an individual, a "netizen" …[with the] means of expression for anyone with minimal technical skills but abundant passion and dedication" (Harden 1998). The Guardian newspaper described the Museum of Online Museums (http://coudal.com/moom/) thus: "The internet has brought advancements, but nowhere has it been more successful than in the field of meaningless rubbish. Here, vast swathes of tat are housed in one handy place for easy navigation" (2007, 31). Just because the creator describes their collection as a museum does not mean to say it functions as we expect of a memory institution, whatever that may be.

However, the content of these online sources ranges from the amusing, to serious attempts at providing information resources to both scholarly and amateur researchers which are just not available anywhere else, being useful even if they lack the institutional backing and guidance of their official online counterparts. These "museums" can vary from the ramshackle and quirky to the glossy and guidelines-compliant documentation of ephemera which established institutions are either not interested, able, or willing to catalogue, digitise, and provide online: "one librarian's ephemeron is another's invaluable cultural resource" (Gorman 2003, p. 14). The Museum of Online Museums maintains a registry of such creations, including Devil's Rope: The Barbed Wire Museum (http://www.barbwiremuseum.com/index.htm), the Museum of Menstruation and Women's Health (www.mum.org), and Total Rewind, "the virtual museum of vintage vcrs" (http://www.totalrewind.org/): all award winning, and featuring exhaustive documentation and digitised source material not available anywhere else. Additionally, many amateur digitisers are creating "pools" of digitised objects utilising image-hosting sites such as www.flickr.com as a platform, creating exhaustive documentation of, say, vintage dressmaking patterns (http://www.flickr.com/groups/vintagepatterns/pool/), or book cover artwork of cheap paperbacks from the mid twentieth century (http://www.flickr.com/groups/paperbacks/pool/).

Memory institutions themselves are beginning to experiment with Web 2.0 environments, asking the general public to interact with their digitised material through social tagging, bookmarking, and commenting (http://www.steve.museum/). A forward-thinking project at Oxford University, the First World War Digital Poetry Archive (http://www.oucs.ox.ac.uk/ww1lit/), has taken this one step further by successfully asking the general public to come forward with their ephemera to include in the archive. Harnessing the energy, passion and interest of amateur digitisation is of clear interest to the cultural and heritage sector. What do we know about both the creators and users of amateur digitisation projects? What can we learn from this?

There has never been an over-arching academic consideration of amateur digitisation projects. This paper surveys the hitherto ignored phenomenon of virtual and online museums and digitised ephemera created by amateur enthusiasts, to ascertain the motivation, scope, implementation, perception, and usefulness of such activity. Is this predominantly meaningless "tat", or are virtual collections created by amateurs used, useful, and worthy of further consideration? By what criteria can we judge whether an amateur digitisation project is successful? How can memory institutions harness the energy and time devoted in creating these online resources?

Methodology

First, the literature on digitisation was reviewed to ascertain whether amateur contributions had been studied. Second, a hundred stand alone, self-confessed "virtual museums" were reviewed to indicate the coverage, scope, and purpose of their collections. Likewise, groups and pools on flickr were reviewed. Memory institutions currently encouraging user interaction via Web 2.0 technologies were surveyed to ascertain the extent of user involvement. Third, ten creators of amateur websites were interviewed to gain their insight regarding purpose, coverage and use of their material. Finally, a survey was carried out with Arts and Humanities academics, to ask if they had every used, referenced, depended on, or even come across useful online digitised material provided beyond institutional boundaries.

Findings

The study will report fully in Spring 2009, but preliminary findings indicate that successful standalone virtual amateur museums – those providing novel detailed content unavailable elsewhere – tend to focus around a particular niche subject such as histories of specific technologies, or socially taboo interests. Another popular area is the digitisation of family history and genealogical material. Those utilising image hosting facilities, such as flickr, unsurprisingly tend to focus on image based material to facilitate discussion of the history of graphic design and illustration.

The digitisation is carried out as a not-for-profit hobby: the interaction with other enthusiasts and viewers afforded by using Internet technologies gives a sense of camaraderie and often encourages rigorous debate between enthusiasts keen on properly documenting their chosen topic. This enthusiasm is carried over to established memory institutions which offer amateurs the means to contribute via web 2.0 technologies.

There has been very little investigation or understanding of how these amateur digitised collections are used. Creators are generally aware of usage statistics, and most can provide examples where specific, detailed queries from interested researchers have been answered through their collections. Academic researchers are happy to turn to these collections when they provide information not available elsewhere.

Conclusion

Enthusiastic digitisation by amateurs, a phenomenon previously ignored by information professionals, is providing a rich source of online cultural heritage content which often documents areas not covered via traditional institutions. The energy and zeal displayed by amateur digitisers is worthy of further consideration, as amateur collections often complement existing collections, providing an alternative free discussion space for enthusiasts. Web 2.0 technologies present great potential in linking the amateur with the institution, extending the reach and scope of digitised cultural heritage.

References

Deegan, M. and Tanner, S. (2002), *Digital Futures: Strategies for the Information Age*. Digital Futures Series (London, Library Association Publishing).

Gorman, M. (2003), "Cataloguing in an Electronic Age", in Intner, S. S., Tseng, S. C., Larsgaard, M. L. (eds.) (2003), *Electronic Cataloguing: AACR2 and Metadata for Serials and Monographs*, (Binghamton, The Haworth Press), 5- 17.

Guardian (2007), "Web Watch", *The Guardian,* Guide Magazine, October 20 2007, 31.

Harden, M. (1998). "Web Graphics: Art on the Net". Museums and the Web Conference 1998, Toronto, Ontario, Canada, April 22-25, 1998. http://www.archimuse.com/mw98/papers/harden/harden_paper.html

Hughes, L. (2004), *Digitizing collections: strategic issues for the information manager* (London: Facet Publishing).

Lee, S. (2002), *Digital Imaging, a Practical Handbook* (London, Facet Publishing).

Minerva (2003), "Summary of Progress". *Coordinating digitisation in Europe*. Progress report of the National Representatives Group, European Commission, The Information Society Directorate-General, <http://www.minervaeurope.org/publications/globalreport/globalrep2002.htm>.

Interactive Visual Analysis of Personal Names in Japanese Historical Diary

Alejandro Toledo
Ritsumeikan University
alex@ice.ci.ritsumei.ac.jp

Ruck Thawonmas
Ritsumeikan University
ruck@ci.ritsumei.ac.jp

Akira Maeda
Ritsumeikan University
amaeda@media.ritsumei.ac.jp

Fuminori Kimura
Ritsumeikan University
fkimura@is.ritsumei.ac.jp

Historical diaries provide information about and facilitate understanding of daily life during their periods. The historical diary of an aristocrat not only contains historic facts but might also help us disclose them. In this paper, we present an interactive web-based system for visualizing aristocrat names mentioned in a historical Japanese diary called "Hyohanki" written by an aristocrat during the late Heian era (1132-1184). In our web-based system, the stacked graph is utilized to dynamically analyze the time-series of those aristocrat names. We have found that trends in the name occurrence and co-occurrence visualized by the system correlate well with historic facts regarding the rise and fall of power of as well as the confrontation among the corresponding aristocrats.

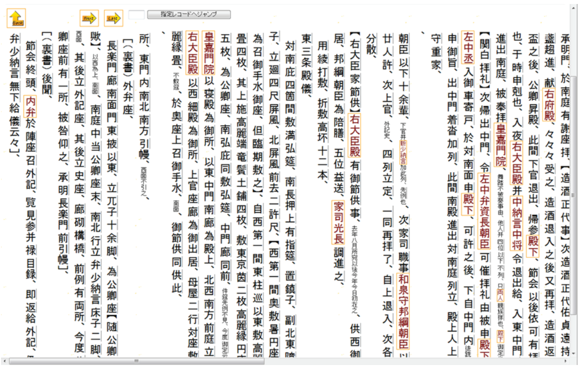
1. Introduction

A diary is a daily record of events that have happened over the course of a day. Diaries written long time ago or historical diaries provide information about daily life during their periods. They can be also utilized to facilitate understanding of the life and times of mentioned individuals in those days. Historical diaries written by aristocrats not only contain historic facts but might also help us disclose such facts.

Information visualization techniques have been successfully applied to historical artifacts. Typical applications include Picasso's artworks and documents (Audenaert et al., 2008; Meneses et al., 2008), 17th-century Portuguese shipbuilding treatises (Furuta et al., 2007), and Vertot's Roman Revolutions (Jensen, 2006). In this paper, we apply information visualization to a historical diary called "Hyohanki (or Heihanki)" which is a diary written by a Japanese aristocrat, Taira no Nobunori, during the late Heian era (1132-1184).

Hyohanki is a relevant resource for the research of Japanese culture of that time period. Although some part of Hyohanki has been deteriorated and missing, all remaining pages are digitized into the text format (Fig. 1), awaiting for digital humanities research activities. An example of such digital humanities research is Cross-Age and Cross-Cultural Information Retrieval discussed in (Maeda and Kimura, 2008).

In this paper, we present an interactive web-based system[1] for visualizing aristocrat names mentioned in Hyohanki. In our system, the stacked graph (Wattenberg, 2005) is utilized to dynamically analyze the time-series of those aristocrat names. This kind of dynamic analysis helps the viewer to track trends in the name occurrence and co-occurrence and thus facilitates revealing of historic facts behind them.



承明門、於南庭有謝座拝、【造酒正代事】次造酒正代佐貞遠持
盞趨進、献右府殿、々々受之、造酒退入之後又再拝、造酒返
盃之後、公卿昇殿、此間下官退出、帰参殿下、節会以後依可有拝
也、于時申剋也、入夜右大臣殿并中納言中将令退出給、入東中門、
進出南庭、被奉拝皇嘉門院、舞踏不被奏事由、他人并四位以下不列、只両人親族拝也、殿下御定
【関白拝礼】次帰出中門、令左中弁資長朝臣可催拝礼由被申殿下
左中丞入御車寄戸、於対南面申殿下、可許之後、下自中門内徒跣
申御旨、出中門着沓加列、此間南殿進出対南庭列立、殿上人上
守重家、
、
朝臣以下十余輩、下官并新少納言加此列、先例也、次家司職事和泉守邦綱朝臣以
廿人許、次上官、外記史、四列立定、一同再拝了、自上退入、次各
分散、
【右大臣家節供】右大臣殿有御節供事、去年八月所宛以後今年今日初在之、供西御
居、邦綱朝臣為陪膳、五位益送、家司光長調進之、
用綾打敷、折敷高坏十二本、
東三条殿儀、
対南庇四箇間敷満弘筵、南長押上有指筵、置鎮子、副北東障
子、立廻四尺屏風、北屏風前去二許尺、【西第一間奥敷暑円座
為召御手水御座、但臨期敷之】、自西第一間東柱巡以東敷高麗
畳四枚、其上施高麗端竜鬢土鋪四枚、敷東京茵二枚高麗縁円
五枚、為公卿座、南弘庇同敷弘筵、中門廊同前、件筵先例不見、今度御定云
皇嘉門院以寝殿為御所、以東中門南廊為殿上、北西南方前庭立
右大臣殿以西細殿為御所、上官座廊為御出居、母屋二行対座敷
麗縁畳、不敷筵、於奥座上召御手水、南面、御節供同供此、
、
所、東門内南北南方引幔、西面不引之、
【(裏書)外弁座、
長楽門廊南面門東掖以東、立兀子十余脚、為公卿座【随公卿
歟】、以西為上、南面、南庭中当公卿座末、南北行立弁少納言床子二脚
西面、其後立外記座、其後立史座、廊砌構橋、前例有両所、今度
卿座前有一所、被咎仰之、承明長楽門前引幔】、
【(裏書)後聞、
節会終頭、内弁於陣座召外記、覧見参并禄目録、即返給外記、
弁少納言無下給儀云々】、

Fig. 1 Example of a digitized page of historical Japanese diary "Hyohanki"

2. Methodology

2.1 Data

Our system is based on entries derived from the Hyohanki Diary. Such entries, contained in a delimited text file, are expressed by a tabular representation consisting of three attributes: (i) year, (ii) aristocrat name for which the family name is placed before the given name, and (iii) a numerical value indicating the number of occurrences of that name in that year. This data was extracted from "Personal Names Index" of Hyohanki, which was manually created by experts in Japanese history.

2.2 Visualization

The method used to visualize the data is straightforward: given a set of aristocrat-names time series, a set of stacked graphs is produced, as shown in Fig. 2. The x axis corresponds to year and the y axis to occurrence ratio, in percentage, for all names currently in view. Each stripe represents a name, and the width of the stripe is proportional to the ratio of that name mentioned in a given year. The stripes are colored blue, and the brightness of each stripe varies according to the number of occurrences, so that the most mentioned names for the whole period are darkest and stand out the most.

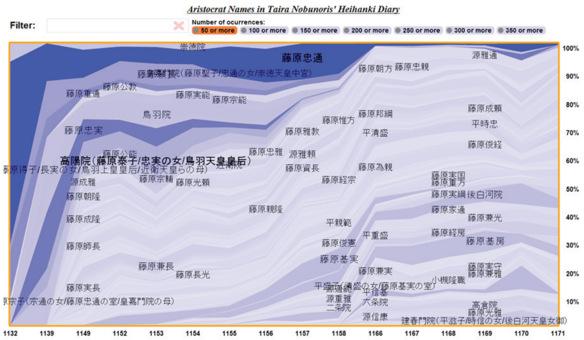


Fig. 2 Screenshot of our Hyohanki's aristocrat-names visualization system, where the English names of those discussed in the paper were manually superimposed under the corresponding Japanese names

Our visualization approach can be seen as an evolved version of timeline representations (Jensen, 2006). The main difference between these two approaches is that timeline representations are suitable to highlight the temporal evolution of events, while stacked graph representations help users in discovering trends in data. In a typical timeline presentation, time is arranged along one dimension and a number of markers, representing events, are placed appropriately along the time dimension. On the other hand, stacked graph representations often create an environment being representative of the data set in question. This environment helps users in discovering trends in data by creating depictions of data values that can make data analysis faster.

2.3 Interaction

When the system starts, the viewer sees a set of stripes representing all names with the number of occurrences above 50. Additional filtering of this data is achieved using two interaction controls. With the first one, *filtering by names*, the viewer may type in letters, forming a prefix; our system will then visualize data on only those names beginning with that prefix. This system reacts directly with each keystroke. Thereby, it is not necessary for the viewer to press return or to click a submit button. In addition, the system moves smoothly between visualization states. So when a letter is typed, an animated transition helps preserve the visualization context.

With the second interaction control, *filtering by number of occurrences*, the viewer can change the data currently in use from the default. As shown in Fig. 2, there are seven buttons, each one allowing the change using the number of occurrences above 50, 100, 150, 200, 250, 300 and 350 respectively. The idea behind this interaction control is that we can restrict the view to certain data of interest, according to their number of occurrences, resulting in concise views of the data.

3. Results and Discussions

Fig. 2 shows the stacked graph of the aristocrat names with the number of occurrences above 50. One prominent stripe can be observed, i.e., that of Fujiwara no Tadamichi (藤原忠通), who was the eldest son of the Japanese regent Fujiwara no Tadazane (藤原忠実). It should be noted that in 1156 the Hogen Rebellion took place between the defeating side of Emperor Go-Shirakawa, sided by Fujiwara no Tadamichi, and the defeated side of retired Emperor Sutoku, sided by Fujiwara no Yorinaga who is the younger brother of Fujiwara no Tadazane.

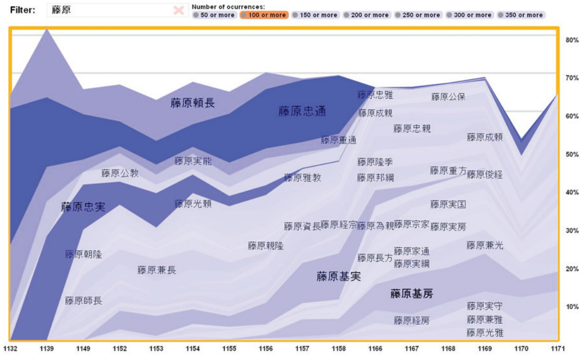

Filter: 藤原

Fig. 3 Stack graphs of names beginning with Fujiwara (藤原), where the English names of those discussed in the paper were manually superimposed under the corresponding Japanese names

Fig. 3 shows the stacked graph of the aristocrats with family name Fujiwara (藤原), a powerful regent family dominating the Japanese politics of Heian period, and with the number of occurrences above 100. It can be seen that the stripe of Fujiwara no Yorinaga (藤原頼長) has high co-occurrence with that of Fujiwara no Tadamichi until the former was defeated by the latter in the aforementioned rebellion. It is also interesting to see that the stripe of their father Fujiwara no Tadazane also conforms to this trend. It should be noted that the y-axis is always the total percentage. This is because we are interested in the relative transition of the influence of a particular family or a person, and for that purpose, the total percentage is better than the relative percentage.

In addition, in Fig. 3, Fujiwara no Motozane (藤原基実) has high occurrence before 1166, but his occurrence decreases after 1166. On the other hand, Fujiwara no Motofusa (藤原基房) has low occurrence before 1166, but has high occurrence after 1166. These also conform to the following historical fact. In 1166, a regime change occurred. Before 1166, Emperor Nijo group was in power. However, Emperor Nijo, the eldest son of the aforementioned Emperor Go-Shirakawa, died in 1165, and Fujiwara no Motozane who was the leader of the emperor Nijo group also died in 1166. Their death triggered the emperor Nijo group to lose their political power. After their death, Emperor Go-shirakawa regained political power. Ultimately, Fujiwara no Motofusa was appointed as the leader of the regime.

4. Conclusions and future work

We have successfully applied the stacked graph to visualization of aristocrat names in the Hyohanki diary. Two interaction controls are provided, i.e., filtering by names and filtering by number of occurrences. They allow the viewer to search names by prefix and to narrow the target names, respectively. Interesting trends have been found that correlate well with the corresponding historic facts.

As our future work, we plan to use data mining techniques to the diary in order to obtain structural representations other than the tabular one used in the current work. In the current version of our system, tabulated data are visualized by means of a time-line that tells us trends of aristocrats names mentioned in the diary. We believe that using data mining techniques will allow us to find new useful information, such as place names, building names, and street names, that eventually leads to different graphical representations.

Notes

[1] http://www.ice.ci.ritsumei.ac.jp/~ruck/IV/Hyohanki.html

References

Audenaert, N., Lucchese, G., Sherrick, G., and Furuta, R. (2008). *CritSpace: Using Spatial Hypertext to Model Visually Complex Documents. Book of abstracts for the DH2008 conference*, Oulu, Finland, pp. 50-53.

Furuta, R., Castro, F., and Monroy, C. (2007). *Ancient Technical Manuscripts: the Case of 17th-century Portuguese Shipbuilding Treatises*. Book of abstracts for the DH2007 conference, University of Illinois, USA, pp. 67-69.

Jensen, M. (2006). Semantic Timeline Tools for *History and Criticism. Book of abstracts for the DH2006 conference*, Sorbonne, Paris, pp. 67-69.

Maeda, A. and Kimura, F. (2008). *An Approach to Cross-Age and Cross-Cultural Information Access for Digital Humanities. Digital Resources for the Humanities and Arts 2008 Conference (DRHA08),* Cambridge, U.K., Sep.

Meneses, L., Furuta, R., Mallen, E. (2008). *Exploring the Biography and Artworks of Picasso with Interactive Calendars and Timelines. Book of abstracts for the DH2008 conference*, Oulu, Finland, pp. 160-163.

Wattenberg, M. (2005). *Baby names, visualization, and social data analysis. Proceeding of IEEE Symposium on Information Visualization 2005* (*InfoVis2005*), pp. 1–7.

Integrating Images and Text with Common Data and Metadata Standards in the Archimedes Palimpsest

Doug Emery
Emery IT
doug@emeryit.com

Michael B. Toth
Walters Art Museum/R. B. Toth Associates
mbt.rbtoth@gmail.com

This paper by the Archimedes Palimpsest program and data managers will discuss the issues, challenges and decisions related to the integration of all images and transcriptions of the Archimedes Palimpsest in digital form. This will include discussion of the management of image and transcription data, adoption of metadata and text encoding standards and schemas, and challenges faced in integrating these data during this 10-year long program and use in follow-on efforts.

This paper will address the integration of scientific and scholarly data through the application of best practices in standardized metadata to images and XML transcriptions, and challenges encountered in applying the various imaging, identification and encoding standards. This will include discussion of the following:

1. The effective implementation of broadly accepted metadata models and data architectures;
2. Integration of integrated data standards, including Dublin Core and Text Encoding Initiative; and
3. Embedding metadata elements within the data itself for effective preservation and archiving, as well as spatial linkage of data elements.

On 22 October, 1998, the Archimedes Palimpsest was sold at auction for $2.0 million to an anonymous buyer. A multidisciplinary team of conservators, imaging scientists, scholars, and information technology professionals disbound, conserved, digitally imaged, analyzed and transcribed the 184 parchment pages for continued study. The program applied advanced spectral imaging to study the Archimedes and other significant medieval manuscripts from the 10th century that were copied over by 13th-century prayer book text. On October 29, 2008, the approximately 1 Terabyte of Archimedes Palimpsest integrated image and transcription data were released to the public for free use. Integrating the ancient Greek transcriptions of Archimedes' mathematical texts with digital images and hosting them on the Web for a broad set of global users posed a complex set of information sharing challenges.

The Archimedes Palimpsest Digital Product required the integration of imaging, scholarly and data products. The product incorporates registered images for each leaf linked spatially to diplomatic transcriptions that scholars initially created in various nonstandard formats, with associated standardized metadata. Imaging scientists included Dr. Roger Easton Jr. from the Rochester Institute of Technology, Dr. Keith Knox from Boeing Corporation, and Dr. Bill Christens-Barry from Equipoise Imaging, and a camera provided by Stokes Imaging, supported by John R. Stokes. The imaging effort built on imaging of the Dead Sea Scrolls by a team from RIT and the British National Gallery Vasari Project, while the transcription encoding effort built on the work of the Homer Multitext Project by the Center for Hellenic Studies. With over 4,000 digital images in 12 spectral bands and 140 pages of transcriptions of the original writings in Greek of Archimedes and Hyperides, standardized metadata was critical to 1) access to and integration of images for digital processing and enhancement, 2) management of transcriptions from those images, and 3) linkage of the images with the transcriptions.

This effort produced images and transcriptions of the only copies of Archimedes treatises The Method and Stomachion; the only copy in Greek of On Floating Bodies; and copies of the Equilibrium of Planes, Spiral Lines, The Measurement of the Circle, and Sphere and Cylinder. It also discovered ten pages of text by the fourth century B.C. Attic Greek orator Hyperides; six folios from a still unidentified Neo-Platonic philosophical text that may be commentaries on Aristotle; four folios from a liturgical book; and twelve pages from two other books, the text of which has yet to be deciphered. These texts are being studied by scholars from a range of colleges and universities, including Oxford, Cambridge, Stanford, and Eötvös Loránd (Budapest) Universities. These scholars bring not only the knowledge and ability to read the sometimes almost illegible ancient Greek text and diagrams, but also significant knowledge of the science, mathematics, law and philosophy discussed in the texts. Capturing this data from a range of scholars and rendering it in a common, standardized digital format required establishment of a program specification for transcribed text.

Beginning in 2001, the Archimedes Palimpsest program adopted established metadata standards to ensure key parameters were recorded during technical collection for use in subsequent processing and studies, including Dublin Core and Text Encoding Initiative guidelines. These standards have been further refined and adapted to address the needs of scholars, imaging scientists, conservation and preservation professionals, and information and data managers. Working with the scholars, the program developed a Transcription Integration Plan for the Archimedes Palimpsest Program that incorporated the Unicode, Dublin Core and Text Encoding Initiative Standards and Guidelines, which proved essential to the integration of the transcribed information. The selection of broadly accepted and up-to-date international consensus standards is an effort to ensure currency, increase the likelihood of long-term data viability, and provide for ample documentation to describe the bit structure of all archive components, from the core data to the supporting files. The program also developed a system architecture that was documented with archival, metadata and integration plans and implemented after extensive review and modifications.

The construction of the data set addresses the special problem of building an archive for today and the distant future. A guiding principle of the archive is the integration of data and metadata components, following principles described in the Consultative Committee for Space Data Systems (CCSDS) *Reference Model for an Open Archival Information System (OAIS)*. Each image bears embedded identifying, spatial, scientific, format, and content metadata in its header. Each directory contains all images for a given folio side, accompanying XMP metadata files, checksum data, and spatially mapped TEI XML transcriptions for the Archimedes and Hyperides texts. Each image file or folio directory forms a self-contained unit of data and preservation information. The directory as a whole provides files that guide users to the data and document the data set and the technologies comprised in it. A simple, documented archive structure supports the discoverability and accessibility of the data.

The archive, the transcriptions, and supporting metadata are designed to support the core image data using broadly accepted standards. The key to the processing and presentation of the Archimedes image data is the registration of all the images of a single folio side to one another. These relationships are documented in and exploited by the supporting files and metadata. The project-developed Archimedes Palimpsest Metadata Standard (APMS) provides a metadata structure specifically geared to relating all images of a folio side in a single multi- or hyper-spectral data “cube.” It relates the components of this cube to the imaged object, the conditions and systems used in its imaging, the standards and techniques used to generate the digital file, and finally the standards used to document this components. Each image has embedded in it its own metadata and so may stand alone or be related to any or all of the other members of the same cube. The standard is based on the Dublin Core metadata elements and Federal Geographic Data Committee’s (FGDC) Content Standard for Digital Geospatial Metadata. The included transcriptions, written in compliance with TEI release P5, support the images and serve as a kind of metadata. For the majority, 142 of 180 folio sides, each is provided with a transcription. In those transcriptions, each line is mapped to a rectangular region of the related images. The TEI `<facsimile>` element and its children are used for this purpose. These digital transcriptions provide a machine-readable tool that document the content of the images. The spatial mapping allows easy mapping from transcription to image and vice versa.

The use of standardized data sets allows the hosting and integration of image and textual data, as well as data from other cultural works, across a range of services providers, libraries and cultural and educational institutions, and the separate development of graphical user interfaces (GUIs) by users as needed. Inclusion of standards as part of the data set will help ensure the data will be readily searchable, available and accessible for studies in decades to come.

Additional information on this program is available at the Archimedes Palimpsest website

References

Dublin Core Metadata Initiative, *Dublin Core Metadata Element Set*, Version 1.1., 14 Jan., 2008.

Platform Models for Scholarly Journal Publishing: A Survey and Case Study

Sarah Toton
Emory University
stoton@emory.edu

The past six years have seen a proliferation of online scholarly journals. The number of new publications added to the Directory of Open Access Journals (DOAJ), for example, has gone from a trickle of twenty-six new journals in 2002 to a torrent of five hundred ninety-eight new titles indexed in the first nine months of 2008. [1] This proliferation has led to publications focusing on more specialized topics, and publishing in nations and languages traditionally underserved by Western-focused academic publishers. Despite the wide variety of emerging content, however, the technical infrastructures behind these publications remains relatively uniform. Constrained by current platform options and multimedia literacy on the part of publishers, many journals still operate akin to their print-based predecessors. By offering electronic text-driven arguments, online journals may challenge conventional models of delivering scholarly content, but do little to augment the emerging shape of digital scholarship. In short, a downloadable PDF is not that different than a scanned photocopy.

This paper examines technologies behind current open-access journals, and it evaluates how journal content, preservation, information architecture, and the review process influence and are influenced by available publishing frameworks. To do this, I survey platforms used currently in open-access scholarly publishing. I then focus on a particular journal, *Southern Spaces*, and its transition from hardcoded HTML to a publishing/content management system. Through this discussion, I intend to address several questions, including: Can journals impact/build scholarly communities? Can they shape and enhance a participatory culture on the scholarly level? Are existing models of peer-review unnecessarily limited by restrictive platforms? Is the text-based model of scholarly communication good enough for today's scholar?

The paper's survey portion examines existing models used in electronic publishing from XML schemas (like NCBI's Journal Publishing Tag Set) to the many open-source platforms developed at digital scholarship centers and research libraries, including Open Journal Systems (OJS) and Digital Publishing System (DPubS). The Public Knowledge Project's OJS platform facilitates the development of open-access scholarship by not only offering an infrastructure for the online presentation of articles, but also providing a management system for peer-review and general editorial workflow. A local install, ease of configuration, and submission management tools allow users to develop a technical infrastructure relatively quickly and with little need for system administrator support in the maintenance phase. The relatively uniform look of OJS journals, however, suggests little capability for extensive customization. Cornell and Penn State's DPubS software offers more opportunities for customization through its modular architecture, as well as the potential for interoperability with Fedora and DSpace repositories. In addition, like OJS, DPubS 2.1 offers a service for peer-review management.[2] While DPubS robust design allows for more unique journal instances, however, it also requires significantly more back-end management and support than OJS. This makes DPubS useful on an institutional level in a library, but the steep learning curve and technical requirements may hinder adoption among university faculty. Other systems developed in Europe—Hyperjournal, SOPS, the ePublishing Toolkit—seek to provide personal archives, workflow support, publishing networks, and cross-referencing tools, but have not yet been adopted by more than a handful of scholarly publishers.

In addition to offering a survey of platform-based publishing options, this paper examines open-source blog publishing tools and content management systems developed for commercial purposes that are also used by electronic scholarly publishers. *Flow,* a media studies journal, and the *CodeForLib Journal*, for example, uses the popular blogging platform, WordPress.[3] *Museum Anthropology Review* used WordPress in 2007 before switching to OJS in 2008. The journal's editors still maintain the WordPress site as supplementary weblog to *MAR*.[4] While implemented less commonly, the content management system, Joomla, provides the platform for Boston College Law School's *Intellectual Property and Technology Forum*.[5] These cases address why publishers chose and sometimes abandoned non-academic platforms, and they offer insight into the possibilities and drawbacks in modifying commercial-based open-source tools.

This comparative survey illustrates the varied strengths and weaknesses of particular publishing platforms, as well as examines their influence on the final product: the published scholarship. I then turn to *Southern Spaces* (www.southernspaces.org), an open-access, interdisciplinary journal housed in Emory University's Woodruff Library. My experience as an editor, media coordinator and the Managing Editor has shown me how technological decisions influence the daily operations of journal publishing, as well as the final shape of scholarly content. This year, in particular, marks a major transition for the journal. *Southern Spaces* will undergo a substantial redesign in the coming months, transitioning from HTML pages created in Dreamweaver to a Drupal 6 platform integrated with a Fedora repository.

To determine the new information architecture for Southern Spaces, I conducted two surveys: one to board members and peer reviewers, the other to junior and independent scholars new to *Southern Spaces*. Based on user feedback as well as an internal we conceived a new workflow process, navigation structures, pedagogical tools, as well as several opportunities for community development within the site. The redesign, I anticipate, will allow *Southern Spaces* authors and staff to develop multimedia publications that incorporate links, video, audio, and Flash features. In addition, the process will develop interactive features that not only allow for reader feedback, but also provide a better infrastructure for citing and collecting *Southern Spaces* articles. These directions extend from *Southern Spaces*'s goals, which are likely similar to many open-access journals: offer a research tool, expand the visitor community, sustain and preserve digital content, streamline the pre-publication process, develop a learning environment for interdisciplinary scholars as well as scholars new to digital scholarship.

Choosing a publishing platform not only dramatically influences published content, but also has the potential to change the landscape of online scholarly communication. Who edits, reviews, publishes and hosts a journal influences technological decisions, but these decisions also dramatically impact what can and cannot be displayed as vetted scholarship, as well as the limits of scholarly participation in a journal's (or a discipline's) wider community. In this rapidly expanding landscape of online scholarship, technological change and the decisions behind this change aren't always apparent. Yet, in order to understand how digital publishing is changing, it is key to look beyond the number of online journals to the variety of online publishing options and their impact beyond the computer screen.

References

Borgman, C. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: The MIT Press.

Brown, A. (2002). XSD Schemas in Book and Journal Publishing. *XML Europe*. http://www.idealliance.org/papers/xmle02/dx_xmle02/papers/03-01-02/03-01-02.html (accessed 11 November 2008).

Sparc Europe (2008). Open Access Journals: Overview. http://www.sparceurope.org/resources/hot-topics/open-journals (accessed 11 November 2008).

Solomon D. (2008). *Developing Open Access Journals: A Practical Guide*. Oxford: Chandos Press.

Willinsky, J. (2005). Open Journal Systems: An Example of Open Source Software for Journal Management and Publishing. *Library Hi-Tech* 23: 504-519.

Notes

[1] Gavin Baker (2008). Growth of DOAJ: Steady 2003-2007, Major Spike in 2008. A Journal of Insignificant Inquiry. October 17, 2008. This study is somewhat limited as the DOAJ offers only two pieces of chonodata: date journal was started and date journal was added to the DOAJ. This study indicates when journals were added to the DOAJ, but not necessarily when became open access.

[2] This build was launched in June 2007, and it remains the most recent version as of November 2008. Documentation for 2.1 was last update in March 2008.

[3] The *CodeforLib Journal* (http://journal.code4lib.org/) currently uses WordPress 2.6.3. *Flow* uses 2.5 (for their second redesign, Flow used a WordPress skin called The Morning After, which gave the publication a magazine style; they currently use the MimboPro template, also designed for magazine style sites.) In late 2008, a DOAJ Export Plugin was developed for WordPress, allowing publishers of OA journals to provide article-level data to the DOAJ, thereby opening articles up to discovery through DOAJ's interface.

[4] The *Museum Anthropology Weblog*'s "About" page explains: "While, during its first year (2007) the journal itself was published on this site, the journal is now published using Open Journal Systems by the Indiana University Bloomington Libraries." http://museumanthropology.net/about/ (accessed 10 November 2008).

[5] Paul Ham (2006). Using Joomla for an Online Law Journal. *Intellectual Property and Technology Forum*. Boston College Law School. Boston, MA. July 9, 2006. http://bciptf.org/blog/2006/07/09/using-joomla-for-an-online-law-journal/. (accessed 9 November 2008). Several comments discuss other publishing platforms, including citing OJS.

Patrick Kavanagh's Poetic Wordscapes: GIS, Literature and Ireland, 1922-1949

Dr. Charles Travis
Trinity College
ctravis@tcd.ie

> Somebody is moving across the headlands
> Talking to himself
> A grey thinker.
> *The Seed and the Soil* (1938)

> a road, a mile of kingdom. I am king
> Of banks and stones and every blooming thing.
> *Inniskeen Road: July Evening'* (1935)

This paper presents the use of Geographical Information Systems (GIS) as a means to digitally explore authorial, literary and historical environments from a multi-sensory perspective. The production of *Patrick Kavanagh's Poetic Wordscapes* has involved the digitization and literary mapping of historical 1902 Ordnance Survey maps of Inniskeen parish in County Monaghan, Ireland, and the townlands of Mucker and Shancoduff. This regional landscape captured in the writings of Irish poet-farmer Patrick Kavanagh (1904-1967) comprised an imaginative hinterland which he observed, inspired many of his works:

> A gap in a hedge, a smooth rock surfacing a narrow lane, a view of a woody meadow, the stream at the junction of four small fields –these are as much as a man can fully experience. As I wander slowly along the over-hanging hedge that separates my fields from the fields of John Woods my past life comes vividly alive in my imagination. Those wonderful days in a world that was only a couple of townlands and yet was eternal.[1]

The GIS visualizations and interactive mappings which comprise *Patrick Kavanagh's Poetic Wordscapes* work are the fruit of an interdisciplinary collaboration between the Digital Humanities, Historical Cartography and Geography and Irish Literary Studies. Humanist geographers have observed that 'literature is the product of perception, or, more simply, *is* perception,'[2] and long maintained a 'view of literature as a valuable storehouse of vivid depictions of landscapes and life.'[3] Correspondingly 'Robert Frost's New England, Gauguin's Tahiti, Hemingway's Spain [. . .] are imaginary places in the original sense of the verb "to image."'[4] Kavanagh's literary landscape of Monaghan was both imagistic and strongly rooted in a farmer's physical relationship with the land itself. As the humanistic geographer Yi-Fu Tuan informs us:

> the entry of nature is no mere metaphor. Muscles and scars bear witness to the physical intimacy of the contact. The farmer's topophilia is compounded of this physical intimacy, or the material dependence and the fact that the land is a repository of memory and sustains hope.[5]

Landscape depictions in Kavanagh's poetry and prose of the 1920s, 1930s and 1940s deeply reflects such a topographical sensibility. In later life Kavanagh observed: 'To know fully even one field or one lane is a lifetime's experience. In the world of poetic experience it is depth that counts, not width.'[6] The vivid images of place in Kavanagh's canon originate from the early memories he possessed of his birth place in the townland of Mucker. One established, they circle outward to encompass the fields of Shancoduff and wider drumlin landscapes beyond, like ripples extending outward upon a pond of poetic imagination. Kavanagh's prose style is particularly suited to GIS visualization and mapping as it inherits the 'Gaelic bardic tradition of *dinnsheanchas* (knowledge of the lore of places)[7] and accordingly reflects an intimate 'geography based on *seanchas*, in which there is no clear distinction between the general principles of topography or direction-finding and the intimate knowledge of particular places.'[8] The base maps utilized in this project are based upon 1902 revisions of surveys conducted by the British Ordnance Survey, which first mapped the island of Ireland between 1824 and 1842 and produced baseline six inch to one mile maps (1:10,560) charting the locations of structures, townlands, fields, roads, streams, as well as Anglicizing Gaelic place names. Whilst topographically accurate, the first ordnance survey in the nineteenth century was a political project of empire and its intent was to create a scale of appraisal, which would establish the basis for Griffith's Valuation of land and property. The 1902 revision maps digitized to visualize the landscapes depicted in Kavanagh's prose are linked to the basic topographical divisions of the 1901 and 1911 censuses: county; district electoral division; townland or street. Original household returns, signed by heads of households, provide a window on the social-economic features of the landscape through three statistical returns, dealing with religious denominations, classification of buildings, and out-offices and farm-steadings. Such an approach has been utilized by Woods and Shelton (1997) in their investigation of the geographical patterns of mortality rates in nineteenth century England and Wales.[9] The mapping revisions published in 1902 and joined with databases containing data from the 1901 and 1911 censuses are commensurate and reflective of boundaries and topographical features described by Kavanagh's account of the social landscape in the first two and half decades of the twentieth century. Utilizing overlayering techniques, these attributes in a GIS system have been geo-coded to features such as boundary lines and structural points, to reconstruct the historical dimensions of particular periods and locations upon the visualized landscape. Digitized maps for *Kavanagh's Wordscapes* were 'rubbersheeted' and draped over a digital terrain model (DEM) created from contemporary geological survey satellite images of the region. Such an approach in digital mapping recreates the poet's visual perspective and movements over a rural drumlin landscape. Selections from Kavanagh's poetry and prose have been geocoded in real-time, through GPS technology in text and audio from point to point of inspiration and provide a 3D model, which recreates his poetic performance through space, as he farmed his fields, walked the winding roads of his townland, and participated in the day-to-day life of his Parish of Inniskeen. A similar technique allowed Harris (200) to employ moving imagery to research the loci of 'sacred space' associated with ancient native American burial grounds in Moundsville, West Virginia.[10] The digital interactive maps comprising *Kavanagh's Wordscape* are but one element of an online literary geography which explores the intrinsic relationship between Irish literature and place during the first half of the twentieth century, and are currently being employed in an ongoing digital humanities project, entitled *A Digital Literary Atlas of Ireland, 1922-1949* to visualize the historic, imaginative and literary landscapes of a selections of Irish authors, by integrating their performances as writers in time and space, with the narratological maps provided by their texts. Though many web based portals in the digital humanities consist of online databases, directories, or repositories of scanned text, this project's nexus between Irish Studies and Historical GIS is thematically different and methodologically a strong one,[11] because GIS's ability to visualize and store attribute, spatial and temporal information, allows the integration of visual, textual and numerical data related to Irish Studies within a spatial frame of reference. This project provides a means for users to visualize the unique relationship between writers, their works and the influence of place, history and culture upon Ireland's literary production in the early twentieth century. The project question being posed, presented and facilitated by GIS is to undercover through the aegis of writers, their locales and their works the heterogeneous nature of Irish identity and its relation to place during the early twentieth century. As Irish poet and Nobel prize winner Seamus Heaney has observed:

> The usual assumption, when we speak of writers and place, is that the writer stands in some directly expressive or interpretative relationship to the milieu. He or she becomes a voice of the spirit of the region. The writing is infused with the atmosphere, physical and emotional, or a certain landscape or seascape, and while the writer's immediate purpose may not have any direct bearing upon the regional or national background, the background is sensed as a distinctive element in the work.[12]

Each selected writer's regional landscape offers a window into the plurality of Irish culture, whether or not this was reflected politically during the period in question. The underlying aim of the project is to provide a re-examination of disciplinary conventions and orthodoxies, provoke research questions and foster public and academic discourse.

Notes

[1] Patrick Kavanagh, *Patrick Kavanagh: Man and Poet,* (ed.) Peter Kavanagh, (Maine: National Poetry Foundation, University of Maine at Orono, 1986) p. 15

[2] Douglas C. Pocock, Humanistic *Geography and Literature: Essays on the Experience of Place*, (New Jersey: Barnes & Noble Books, 1981) p. 15.

[3] D. W. Meinig, 'Geography as an art', *Trans. Inst. Br. Geogr. N.S.* (1983) 316.

[4] Marwyn S. Samuels, 'The Biography of Landscape: cause and Culpability', D.W. Meinig, (Ed.) *The Interpretation of Ordinary Landscapes* (NY/Oxford: Oxford University Press, 1979) p. 70.

[5] Y.F. Tuan, *Topophilia: A Study of Environmental Perception, Attitudes, and Values*, (Englewood Cliffs: Prentice-Hall, 1974) p. 97.

[6] Patrick Kavanagh, *Patrick Kavanagh*, 15.

[7] Declan Kiberd, *Inventing Ireland,* (London: Vintage, 1997) p. 107.

[8] Charles Bowen, 'A Historical Inventory of the *Dindshenchas,*' in *Studia Celtica* 10/11, (1975/76) p. 115.

[9] R. Woods and N. Shelton, *An Atlas of Victorian Mortality* (Liverpool: Liverpool University Press, 1997)

[10] T. M. Harris, 'Moving GIS: exploring movement in prehistoric cultural landscapes using GIS' in G.R. Lock (ed.) *Beyond the Map: Archaeology and Spatial Technologies*. (Oxford: IOS Press, 2000) pp. 116-23.

[11] An increased spatial emphasis is a recent development in the field of Irish Studies and conversely, historians and other researchers in the humanities have begun to incorporate the visual and analytical technology provided by GIS in their studies. In regards to Irish Studies, Patrick Duffy's *Exploring the History and Heritage of Irish Landscapes* (Four Courts Press: 2007) surveys in part how literary and artistic representations of the environment can provide insights into 'imagined worlds of the past.' Tim Robinson's *Connemara: Listening to the Wind* (Penguin Ireland: 2006) explores environmental, cultural and historical dimensions of this Atlantic county by conducting a series of hermeneutic readings of natural and human history within the context of contemplative fieldwork. Andrew Kincaid's *Postcolonial Dublin: Imperial Legacies and the Built Environment* (University of Minnesota Press: 2006) investigates the relationship between ideology and material culture in the development of twentieth century Irish social and political identity, and Liam Kennedy, Paul S. Ell, E.M. Crawford and L.A. Clarkson's *Mapping The Great Irish Famine: A Survey of the Famine Decades* (Dublin Four Courts Press: 1999) incorporates Historical GIS to analyze and visualize various social landscapes of Ireland during the Famine years of the nineteenth century. The latter title is emblematic of the emerging field of Historical GIS which incorporates the study of history and culture within a spatial perspective. Major research projects in this growing field are illustrated by online portals such as the *The Salem Witch Trials Archive* and *The Valley of the Shadow Project* hosted by the University of Virginia's Center for Digital History, *The China Historical GIS* hosted by the Center for Geographic Analysis at Harvard University, *The Historical Atlas of Canada Online Learning Project* hosted by the Department of Geography at the University of Toronto, and the *Great Britain Historical GIS* hosted by the Department of Geography at Portsmouth University. These projects serve as interactive online portals, which can be accessed by interested scholars, educators and the general public.

[12] Seamus Heaney, the *Place of Writing* (Atlanta, Georgia: Scholars Press, 1989) pgs. 20-21.

Googling Google Books: Integrated use of Fragmentary Information Display in Google Book Preview of Electronic Books

Kirsten C. Uszkalo
Simon Fraser University
kcu2@sfu.ca

Teresa Dobson
University of British Columbia
teresa.dobson@ubc.ca

Stan Ruecker
University of Alberta
sruecker@ualberta.ca

Introduction

In previous articles and panels, the authors of this paper have argued that human/textual interaction is an essential missing aspect of the electronic book reading experience. In order to fulfill the needs of the bibliophile, for whom the touch, feel, portability, and engagement with the paper book is part of the pleasure of reading, the electronic book needs to be a tactile, multiplatform device (Ruecker & Uszkalo 2007, 2008). Our investigation led us to look at the user interface of current single platform electronic reading technologies such Sony PRS 505 and Kindle and the more multifaceted iLiad v2 reader/writer. We argued that e-paper makes these devices more user-friendly and offers the potential for a lower environmental footprint then paper books (de Grancy 2008), but noted that their lack of positive form, function, and feel when coupled with their proprietary software and the cost of e-books remained major stumbling blocks to the overall adoption of electronic reading devices (Sottong 2008, Milliot 2008a, 2008b). We concluded that the numerous multifunction platforms used for electronic reading, such as computer screens, phones, and iPods will keep a single-function electronic book reader as a specialty item.

In this phase, we turn our attention to a fuller consideration of one of those platforms: Google Books. Since Google Books has digitized and displays copyright material, it provides the user with incomplete views of the text. Our research question asks how does the underlying structure of Google Books' display of snippets, incomplete chapters, and limited views of text affect how the digitally efficacious student navigates through these documents and how does it affect her reading and research experience, comprehension, and synthesis of material?

The academic library's widespread adoption of digital resources has created the need to reconsider how students are conducting research. Undertaking university-level research assignments has always been a frustrating skill to acquire for new students. Current critical consensus suggests the necessity of teaching specific independent research strategies to university students to help them navigate and synthesize the massive amount of resources available to them (Holz et al., 2008, Femster & Gray 2008, Polack-Wahl & Anewalt 2006). In ebrary's Global Student E-book Survey, released June 2008, students ranked e-books equivalent with their print counterparts as "trustworthy" and acknowledged their use in research assignments. Research using electronic documents provides the opportunity to negotiate broad spans of text; Lui (2005) argues that "screen-based reading behavior is characterized by more time spent on browsing and scanning, keyword spotting, one-time reading, non-linear reading, and reading more selectively, while less time is spent on in-depth reading, and concentrated reading." In terms of conducting academic research, students can sift through information and collect numerous snippets of information. Many students are already used to engaging with multiple internet applications like audio and video streams, instant messaging, and file sharing to access information inside and outside the educational setting (McGreal & Elliot 2004). Likewise, they use functions like Twitter, texting, FaceBook, RSS feeds, keyword searches and indices to find and absorb small fragments of information in meaningful ways. The challenge emerges in synthesizing these snippets of information outside of their original context. However skilled the student might be in amassing information, the information they gather using Google Books often lacks context within the larger work. The student must synthesize the snippets into their own linear arguments without having read or followed the complete arguments in the texts they are mining for quotes. Eshet-Alakali argues that to create "original academic work with the aid of digital techniques for text reproduction, requires scholars to master a special type of literacy" (98). They call this reproduction literacy, which is "the ability to create a meaningful, authentic, and creative work or interpretation, by integrating existing independent pieces of information," which is a *learned* skill (98).

Although researchers have yet to see empirical evidence that students can translate reproduction literacy into successful academic argumentation, the students perceive they can. As such, the movement towards using Google Books' limited previews to quickly search specific ideas across texts not otherwise on hand appears to be a logical extension of existing information finding practices, acting as a center point between authoritative text and quick nonlinear thinking and research strategies. Students are trained to see books as an authoritative source, and they translate the affordance of the printed book in terms of St. Amant's relationships or properties of relationships and perceived properties into their interaction with Google Books as part of a continued, trusted relationship (1998). In terms of internet efficacy (Eastin & LaRose, 2000), and the positive affordances of online learning (Anderson 2004), the dynamic and fragmentary interaction provided by Google Books' platform seems familiar and would not be seen as detrimental to this group of users. However, users may consider just how much their textual interaction is being controlled and delineated by Google Books and how those limitations affect knowledge gathering and comprehension.

The Canadian Association of Research Libraries Copyright Committee argued in their recent survey, "Task Group on E-Books" that there "is a danger that research libraries are adding e-books to their collections using agreements that significantly reduce users' rights" including textual access and reproduction. Issues of copyright are what create different experiences for readers – providing whole texts or stripping access down to sound bites. According to Google Book Search help, for books which are under copyright, and are not part of their partner programs, users are only able to see "basic information about the book, similar to a card catalog, and, in some cases, a few "snippets" of "sentences of [their] search terms in context." These keyword results allow students to feel critical engagement with texts, providing a singular index across an otherwise flat and non-transferable platform. In this way information is displayed in vastly different amounts, dependant on copyright. On some occasions lines of text appear as little scraps of digital paper, or "snippets"; on other occasions, previews allow readers to view a certain number of pages, or read a random sequence of pages at a time, often missing a few pages in between.

Google Books essentially operates under a shroud of secrecy, despite partnerships with numerous universities such Cornell and Columbia, providing no practical information about its underlying code. Despite this, there a few things we can assume about its digital creation process based on common digitization techniques. It is likely that, as part of the Google Books digitization process, two things are created: the scanned image, which is displayed to the user, and an XML-encoded text file generated by an OCR (Optical Code Recognition) application, which is not seen. This XML-encoded text file contains all of the words in the document, but also stores positional information for each word (where each word appears physically on the page, including its height and width). When users search in Google Books, they are not searching the scanned image, but rather, are searching the XML-encoded text file. Google Books then combines the scanned image with the positional information in the XML-encoded text file, and fakes the text highlighting by drawing a transparent yellow box over the word in the scanned image.

Textual engagement in Google Books is therefore radically altered when the display truncates and limits the amount of text a reader can access. The text is at the same time both dynamic and static, neither simply file nor page, and the reader is neither simply a researcher nor a casual reader. Text display and reader interaction remain in flux, dependent on how the text is being displayed and how much the reader can view and search. The reader might find her experience as constructed in small glimpses, large text chunks, or whole chapters, but cannot know how the text will be displayed until she loads it. Likewise, she cannot know when a blank page will remove part of the text and if subsequent pages will appear. However, the successful keyword hit across an otherwise blanked out book is seen to represent a successful, albeit fragmentary and miniaturized, search.

Conclusions

Digitally savvy university students have begun to use Google Books as a research tool. Although this platform provides fragmentary keyword searches and limited previews of text blocks, which radically alter the user's reading experience, anecdotal evidence suggests students who are sufficiently familiar with short text display communication feel they get enough information from the limited preview to use these texts as sources for research. Although Google Books' textual display has the potential to radically unsettle the reading experience it appears as sufficient for entry-level engagement with students who still see these displays as offering the authority of the book and the ease of use of the electronic text. More investigation of the way in which the Google Books' display facilitates research interactions among university students is needed.

References

Anderson, Terry. "Toward a Theory of Online Learning" *Theory and Practice of Online Learning*. Athabasca: Athabasca UP. 2004, 33-60.

Benson, Denzel E., Wava Haney, Tracy E Ore, Caroline H Persell, Aileen Schulte, James Steele, Idee Winfield. "Digital Technologies and the Scholarship of Teaching and Learning in Sociology." *Teaching Sociology*. 30 (2002). 140-157.

de Grancy, Gerald Senarclens. Diss. "Technical, Ecological and Economic Aspects of Electrophoretic Display Applications" *Institute for Information Systems and Computer Media* (IICM) Graz University of Technology A-8010 Graz, Austria, Europe, March 13, 2008.

Eastin, M., & LaRose, R. (2000). "Internet self-efficacy and the psychology of the digital divide." *Journal of Computer Mediated Communications*, 6(1).

ebrary. *2008 Global Student E-book Survey* Online. Accessed November 11, 2008. <www.ebrary.com/corp/collateral/en/Survey/ebrary_student_survey_2008.pdf >

Eshet-Alakalai, Yoram. "Digital Literacy: A Conceptual Framework for Survival Skills in the Digital Era" *Journal of Educational Multimedia and Hypermedia*.13:1(2004), 93-106.

Femster, Nick and Alexander Gray. "Can great research be taught?: independent research with cross-disciplinary thinking and broader impact" *ACM SIGCSE Bulletin*. 40:1 (2008).

Holz, Hilary J., Anne Applin, Bruria Haberman, Donald Joyce, Helen Purchase, and Catherine Reed December "Research methods in computing: what are they, and how should we teach them?" *ACM SIGCSE Bulletin*. 38:4 (2006)

Lui, Ziming "Reading behavior in the digital environment: Changes in reading behavior over the past ten years" *Journal of Documentation*. 61:6 (2005). 700 – 712

McGreal, R., & Elliott, M. "Technologies of online learning (elearning)." *Theory and Practice of Online Learning*. T. Anderson & F. Elloumi (Eds.). Athabasca, AB: Athabasca University. 115 – 136.

Milliot, Jim. "Report Finds Growing Acceptance of Digital Books." *Publishers Weekly*. February 18, 2008a.

Milliot, Jim. "Sony Adopts EPUB Standard for Reader." *Publishers Weekly*. July 24, 2008b.

Owen, Victoria *et al*. "E-Books in Research Libraries: Issues of Access and Use" Online. Accessed November

11, 2008. <http://www.carl-abrc.ca/projects/copyright/pdf/CARL%20E-Book%20Report-e.doc>

Polack-Wahl, Jennfier A., Karen Anewalt. "Learning strategies and undergraduate research" *SIGCSE '06: Proceedings of the 37th SIGCSE technical symposium on Computer science education*. March 2006, 209-213.

Ruecker, Stan, Kirsten C. Uszkalo. "Binding the Eletronic Book: Design Features for Bibliophiles". *Visible Language*. 2007.

St. Amant, Robert. "Planning and user interface affordances" *Proceedings of the 4th International conference on Intelligent user interfaces*. Los Angeles, California, United States, 1998. 135 – 142.

Sottong, Steven. "The Ellusive E-Book." *American Libraries*. May 2008, 44-50.

Medieval Scribes in Parts of Speech

Comparing the vocabulary of different copies of the same text

Karina van Dalen-Oskam
Huygens Instituut KNAW
karina.van.dalen@huygensinstituut.knaw.nl

Introduction

Van Dalen-Oskam and van Zundert (2007) applied the lexical richness measures Yule's K and Burrows's Delta in an authorship attribution study on the Middle Dutch *Romance of Walewein*, which we know was written by two authors. The researchers used a walking window to find where the second author took over from the first author. The change of authors could be seen in the graphs for Delta for three strata of high-frequency words (1-50, 51-100, 101-150); it showed the best in frequency stratum 101-150. In the highest frequency stratum, however, the break between the two scribes responsible for the manuscript of the text (this break occurred much earlier in the text than the author break) was even clearer.

Following up on this research, the researchers looked into the distribution of parts of speech in the above-mentioned frequency strata (van Dalen-Oskam & van Zundert 2008). When they examined the strata, starting with the highest frequency words (1-50) and going on to frequency stratum 101-150, they found an increasing percentage for the categories 'noun' and 'verb', and a decreasing percentage for 'pronoun' and 'preposition'. This may imply that in the area in which this text showed the clearest difference between scribes (i.e. frequencies 1-50), part of the uniqueness of these scribes lies in their idiosyncratic use (as to frequencies) of pronouns and prepositions. In the area in which *Walewein* showed the clearest difference between authors (frequencies 101-150), part of the uniqueness of the authors lies in their idiosyncratic frequencies of nouns and verbs. In other words: the scribes seem to differ especially in the frequencies of function words, whereas the authors seem to differ especially in the frequencies of content words.

Research questions

Authorship attribution for medieval texts before the age of printing is hampered by several practical problems. First, we often do not know a text's author. However, even when the author is known, almost always his or her original text has not come down to us. All that remain are copies of the text, or copies of copies of copies, which are usually difficult to position in a trustworthy family tree (stemma) of the manuscripts. Thus, we do not have the text of the original author, but only texts resulting from manual copying by scribes – persons who made a copy of a text for their own use or for the use of others (who may or may not have paid for these copies and perhaps had special wishes for their copies). We know that scribes made mistakes, and that they changed spellings and wording according to what they thought fit for their audience. And we know that they sometimes reworked the text or parts thereof. In most cases, however, the original text is clearly recognizable. But when a text shows very many changes, scholars start describing these copies as adaptations, and consider them a new text rather than a copy, and see the scribe not as a copyist but as an author in his or her own right. Thus, the question is: how many changes are needed to call a text not a copy but a new text? All in all, the extent of the influence a scribe had on the text he or she copied has been insufficiently studied.

The *Walewein* results led to new research questions. The first was whether we could confirm the findings in a larger set of texts. If scribes do indeed differ as regards how frequently they use certain parts of speech, from the attribution point of view we could ask whether it may be possible to distinguish scribes from each other. From a literary and philological point of view, we may expect a clearer answer to the question in what ways scribes differed from authors (if they did differ, that is). And from a methodological standpoint, we would want to know which measures yield the best results.

Corpus and method

Not many Middle Dutch texts are available in a substantial number of copies. We chose a work by the Flemish author Jacob van Maerlant: the *Rijmbijbel* ('Rhyming Bible'), which is a translation/adaptation of the Medieval Latin *Historia scholastica* written by Petrus Comestor. Van Maerlant finished this work in 1271, and many fragments and fifteen manuscripts (though not all containing all parts of the text) are handed down to us, dating from ca. 1285 to the end of the fifteenth century. One of these manuscripts is available in a good edition; it is also digitally available lemmatized and tagged for parts of speech. Transcriptions of the other manuscripts had to be made for this research. Because of the length of the texts (almost 35,000 lines), we had to work with samples. We chose 5 samples of 200-240 lines from different parts of the text, and transcribed the parallel texts (if available) from all 15 manuscripts, lemmatized the samples and tagged them for parts of speech. The manuscripts are indicated by the letters A, B, C, D, E, F, G, H, I, J, K, L, M, N and O. We approached the samples as 'bags of words' and decided to compare these bags of words as regards vocabulary (number of unique types and tokens) and frequencies of parts of speech. We differentiated between ten parts of speech:

00 noun (content word)

02 proper name (content word)

10 adjective

20 main verb (content word)

21 copula / auxiliary verb

30 numeral

4* pronouns

- **40** personal pronoun
- **41** demonstrative pronoun
- **42** relative pronoun
- **43** interrogative pronoun
- **44** indefinite pronoun
- **45** possessive pronoun

50 adverb

70 preposition

8* conjunction

80 coordinating conjunction

81 subordinating conjunction

82 comparative conjunction

For each part of speech, in each sample we measured the absolute frequency, the relative frequency, the average of the fifteen samples, the standard deviation, the z-score and the ranking of the manuscript in comparison with the other fourteen manuscripts.

Example

To give an impression of the material, the second line of the 'Judith' episode in all fifteen manuscripts is presented below. Judith was 'van liue wijf van herten man' ('as to her body (a) woman, as to her heart (a) man'):

A	Van liue wijf van herten man
B	van liue wijf van herten man
C	Van liue wijf van herten man.
D	Van liue wijf van herte man
E	van liue wijf van herten man
F	Van liue wijf van herten man
G	Van liue wijf. van herten man...
H	Van lyue **een** wijf van herte **een** man

I	Wijf van liue van herten man
J	Van liue wijf van herten man
K	Van liue wijf van herten man
L	Van liue wijf van herten man
M	Van liue wijf van herte man
N	Van liue wijf van herten man
O	van liue wijf van herten man

Manuscript H has two indefinite pronouns, something that none of the other manuscripts has. Manuscript I has a different order in the first noun phrase. The differences in manuscript H will have an influence on word counts and the frequencies of parts of speech, while the difference in manuscript I will not show up in these counts. The differences among the manuscripts for this line are small. Many other lines show more, and more complex differences.

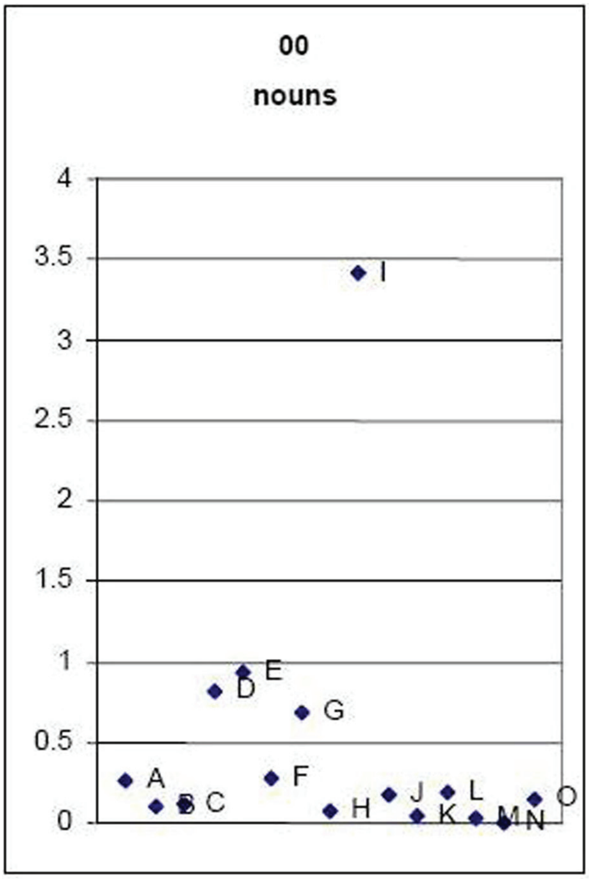


Fig. 1

Results

Reading through all fifteen Judith episodes, we find that two of the manuscripts show remarkable differences compared to the other thirteen. These manuscripts are also the odd ones out in the statistical results. Figure 1 presents the results for the nouns (excluding names). The horizontal axis is the mean of the frequency of nouns in all fifteen Judith episodes. The vertical axis gives the z-scores: the higher a manuscript is located on the graph, the more significant the deviation from the mean. Everything above z-score 1 may be statistically significant. Manuscript I clearly deviates from all the others as to the frequency of nouns. Other parts of speech yield different results (cf. figure 2).

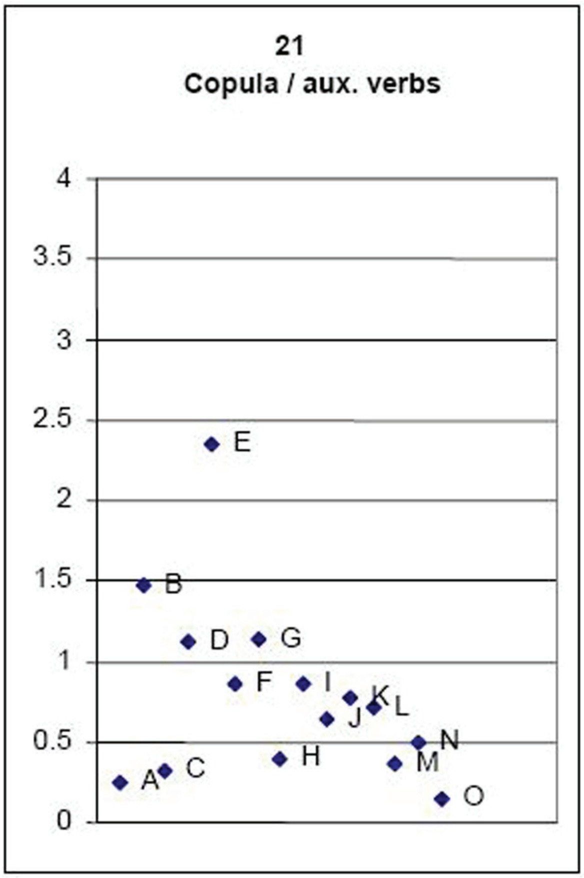


Fig. 2

In the mean of all codes in figure 3, only E and I are above z-score 1. All other manuscripts are below this line. They do, however, show some variation, and from the point of view of the scholar, this variation is interesting. An evaluation of all results for the different parts of speech shows that adjectives differ the most and nouns differ the least. Only manuscript I shows a significant deviation in the frequency of the nouns. Other parts of speech that often differ are pronouns, adverbs and – to a lesser degree – auxiliary verbs/copula. Names, main verbs and conjunctions show more differentiation than the nouns excluding names, but not much. These results seem to confirm that in copying a text, scribes could indeed 'manoeuvre' in the area of function words (adverbs, adjectives, etc.), but may have been kept in check as to content words (nouns, names, main verbs). The only manuscript that deviates wildly in the frequency of nouns (i.e. manuscript I) is special. Reading the text, it is immediately clear to the researcher that, at least in the episode about Judith, it is a very free adaptation of the text. In this episode, the scribe clearly has to be seen as an *author*.

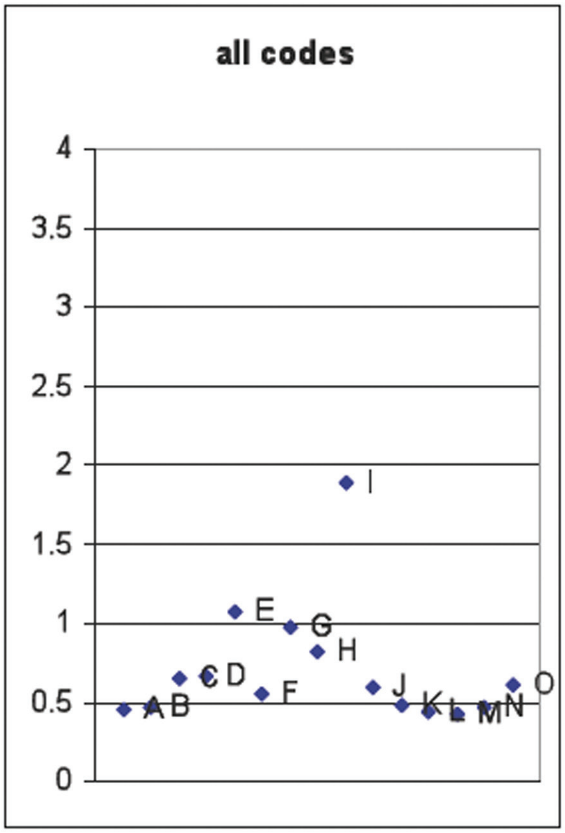


Fig. 3

Our paper will present a comparison with the other samples from the same text. This comparison will show that it is unlikely that we will be able to distinguish scribes with this method – although, as stated, we can clearly spot an 'author' between the scribes. We hope to apply some other measures to the corpus, for example type-token ratio, Yule's K and Burrows's Delta.

Conclusions

The analysis and comparison of the vocabulary of copies of the same text seems to be a promising way to gain more insight into the range of freedoms that scribes were allowed. Of course, both more tests and a much larger corpus are needed. A comparison with other languages would also be interesting. The results of this empirical approach will promote, for example, the development of new tools or the fine-tuning of existing tools for scribal measurements (e.g. Spencer & Howe 2001, 2002).

References

Karina van Dalen-Oskam & Joris van Zundert (2007). 'Delta for Middle Dutch: Author and copyist distinction in "Walewein"'. In: *Literary and Linguistic Computing* 22, pp. 345-362.

Karina van Dalen-Oskam and Joris van Zundert (2008). 'The Quest for Uniqueness: Author and Copyist Distinction in Middle Dutch Arthurian Romances based on Computer-assisted Lexicon Analysis'. In: Marijke Mooijaart and Marijke van der Wal (eds.) *Yesterday's words: contemporary, current and future lexicography.* [Proceedings of the Third International Conference on Historical Lexicography and Lexicology (ICHLL), 21-23 June 2006, Leiden]. Cambridge: Cambridge Scholars Publishing, 2008, pp. 292-304.

Matthew Spencer and Christopher J. Howe (2001). 'Estimating distances between manuscripts based on copying errors'. In: *Literary and Linguistic Computing* 16, pp. 467-484

Matthew Spencer and Christopher J. Howe (2002). 'How accurate were scribes? A mathematical model'. In: *Literary and Linguistic Computing* 17, pp. 311-322

Choreographing the Data: performing the ARTeFACT project TAKE 2

Susan L. Wiesner
University of North Carolina, Greensboro
susan_wiesner@uncg.edu

Jama S. Coartney
University of Virginia
jsc3x@virginia.edu

Rommie L. Stalnaker
Gainesville Ballet Company
rstalnaker81@gmail.com

Argument

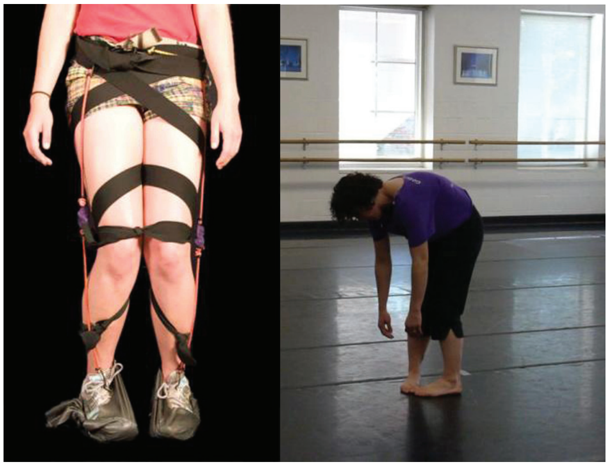
Although one might question the value of collecting and preserving artefacts for all elements surrounding a dance, the preservation of dance as a digitised performance text can become a basis for research that address the issues of describing and analyzing the performance event, while it also enables creativity which in turn further supports research.

Premise

At DH 2008 Susan L. Wiesner and Jama Coartney presented their work on the ARTeFACT project, and posited the previous argument, which resulted in a suggestion (challenge?) to create a work based on the data. As ARTeFACT is a multi-layered, iterative, project involving the contributions of many, the suggested choreography had already been planned as a project step. It is the continued work on this project in both the capture and preservation of movement-driven data (generated from one dance work) and subsequent choreography based on that data that is the focus of this suggested panel/performance.

Method

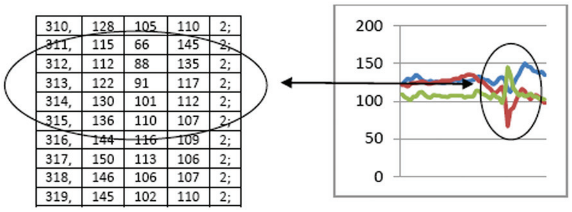
As noted in 2008, the ARTeFACT Project Alpha included teams of students asked to design and build an orthotic device that, when worn, causes the wearer to emulate the challenges of walking with a physical disability. In lieu of a final exam, the students produced a performance during which members of each group wore their devices and moved through the space according to 'choreography' set by their peers. Two cameras preserved the performance itself.



Photos L to R: Ricketts device; choreographic idea; movement abstraction

Inspired by the concept of the devices, a choreographer used the collection of artefacts from the engineering students' course-work to develop a dance work, *For Natalie*, which became ARTeFACT Project Beta. In addition to her study of the digital artefacts generated by the engineering class students, the choreographer used the devices themselves (actually wearing them to choreograph). This forced her to work within the limitations of the devices, inhibiting her level of abstraction, thus causing her to choose movement similar to those of the students, especially in the case of the Ricketts device.

Using Rudolph Laban's Effort/Shape[i] concepts, she also analysed her choreography to create a more cohesive movement vocabulary. Her work was preserved through still photographs, two static and mobile video cameras, and motion capture devices (wiimotes). These wiimotes were placed on the dancers' bodies, and data signals were captured using an interactive authoring program, Max MSP/Jitter. The resultant data, 57,326 samples with 229,304 descriptive elements, was exported as plain text and can be viewed in a custom playback engine or any application that reads plain text files.



310,	128	105	110	2;
311,	115	66	145	2;
312,	112	88	135	2;
313,	122	91	117	2;
314,	130	101	112	2;
315,	136	110	107	2;
316,	144	116	109	2;
317,	150	113	106	2;
318,	146	106	107	2;
319,	145	102	110	2;

Data samples generated by sensor on left calf (from movement abstraction)

From the data, a third work, *For Always*, is being choreographed (completion date December 2008) using Laban's Eukinetics (Effort/Shape) theories. The same preservation techniques will be used against this work in order to find the traces between movement (through established methods of movement analysis such as Laban and Adshead's qualitative approaches) and quantitative analysis of the data samples including the relative placement of the capture devices.

Conclusion

The panel/performance proposed herein considers the processes of iterative choreography and the inclusion of digital technologies into the choreographic process, the performance event, and the preservation of both. It will also discuss the possibilities generated by technologies themselves (towards methods of controlled vocabularies based on movement and/or automated, feature-based tagging). The discussion will culminate in a multi-media dance performance, and an 'after-performance' workshop and Q/A. Thus it is requested that the panel/performance be scheduled for at least 90-minute slot. The panel includes: Jama Coartney, Head of the Digital Media Lab at UVA (originator of the project, who will speak to the technologies and data comparisons); Rommie Stalnaker, Project Beta choreographer (to present her choreographic processes and movement analysis), and Susan L. Wiesner, Project Gamma choreographer (to offer a theoretical analysis of the traces and her choreographic processes).

Sample Bibliography

Adshead-Lansdale, J. (ed.) 1999, *Dancing Texts: intertextuality and interpretation*. London: Dance Books.

Dell, Cecily 1977, *A Primer for Movement Description: Using Effort-Shape and Supplementary Concepts*. New York: Dance Notation Bureau Press.

Goellner, E. W. & Murphy, J. S. 1995, *Bodies of the Text*. New Brunswick, NJ: Rutgers University Press.

Kholief, M., Maly, K. & Shen, S. 2003, 'Event-Based Retrieval from a Digital Library containing Medical Streams' in *Proceedings of the 2003 Joint Conference on Digital Libraries* (Online) Available at http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/8569/27127/01204867.pdf

Naugle, L. 2001,'Reinterpreting Choreography: Motion Capture Data as Historical Information' in *Society of Dance History Scholars Proceedings 2001* 24th Conference, Birmingham: SDHS.

New York Public Library for the Performing Arts, Jerome Robbins Dance Division 2005, Report from Working Group 4 *Dance Documentation Needs Analysis* Meeting 2.

Note

[i] Effort is "how the body concentrates its exertion" (Dell 1977, 10). The elements of Effort are: "direct and indirect, light and strong, quick and sustained, free and bound" (Ibid.). Shape is "how the body forms itself in space" (Dell 1977, 42). Its elements are: "spreading and enclosing, rising and sinking, growing and shrinking, advancing and retreating", as well as movement that goes "sideward out and sideward across, upward and downward, growing and shrinking, forward and backward". Laban's explanation of Shape not only describes movements within the body (the personal Kinesphere), but can also be used to describe movement that traverses the stage space (pathways).

Corpus Analysis and Literary History

Matthew Wilkens
Rice University
wilkens@rice.edu

The problem of periodization has long occupied literary studies. Our ability to distinguish the cultural and aesthetic production of one era from that of another is a basic assumption of our historicizing critical method. Moreover, there exists a broad consensus concerning the general arc of literary history and its major moments. As a practical matter, however, we often find such periods to be both less mutually distinct and less internally uniform than we have been lead to believe. When, for instance, does modernism begin and end? And what, exactly, do Proust, Joyce, Woolf, and Hemingway have in common, to say nothing of mass-market detective and shopgirl fiction from the same era?

Such uncertainty is neither insurmountable nor even especially problematic, but it does emphasize the centrality and the limitations of both close reading and theorization as the working methods of literary study. Because we can read only a finite (and quite small) number of texts, the specific texts that we *do* manage to read will have a disproportionate influence on our understanding of the larger field of cultural production we understand them to represent. Theorization is then frequently dedicated to working back out or up from this restricted corpus of widely read material to the larger set social and economic arrangements that must have been in place so that work of just such a type could have been produced.

There's nothing wrong with this approach, but it would be useful to have relative measures of the extent to which the periods and associated genres we now commonly identify in literary history apply to the full field of literary production over the last several centuries. That is, we would like to be able to answer questions about the coherence of literary periods and genres, their internal variation, their distinction from one another, the sharpness of the breaks between them, and the mechanics of the transitions between them, and we would like to be able to do all this with reference not just to a restricted corpus or canon, but to the broadest possible survey of texts. We would also like to know whether or not there exist discernibly coherent periods, genres, or geographic regions that we have not yet identified, or if our current historical/geographic/generic boundaries are the best possible ones.

The work presented in this paper describes the early steps and initial results of a long-term project designed to address these issues. Building on insights and tools pioneered in corpus linguistics (see especially the work of Mark Davies at BYU), but aiming squarely at questions of literary history and theory, this work begins with the construction of a significant corpus of public-domain literary texts spanning the sixteenth to twentieth centuries from the Gutenberg archive. Gutenberg, of course, is not a typical scholarly resource, but it has the advantage of being large and unencumbered by intellectual property restrictions; one of the intermediate products of the work presented here is an evaluation of its relative merits and suitability in comparison to the smaller but fully curated Wright American fiction and Chadwyck-Healy nineteenth-century fiction collections from the MONK Project. The paper describes the techniques used to construct and to characterize this corpus, as well as the quantitative results of this analysis. It is thus the first full, computationally assisted description and evaluation of the Gutenberg English-language fiction holdings as a resource for digital literary studies.

The paper also advances a set of conclusions based on these results and in dialogue with existing theoretical work on literary-cultural periodization. There is reason to believe that the comparatively brief periods of rapid change between more stable literary eras should be marked by increased incidence of allegorical and tropological language use (this fact of course superimposed on long-term baseline changes in, for example, metaphor generally). Although it is as yet difficult to measure such features directly (but see interestingly related attempts such as Pasanek and Sculley, "Mining millions of metaphors," LLC 23 [2008]: 345-60; Bei Yu's 2006 dissertation on literary text mining; and Matthew Jockers and Franco Moretti's as-yet unpublished work on automatic genre classification), the present analysis suggests that texts drawn from such transitional periods are relatively poor in adjectives and adverbs, a fact that may correspond to their increased reliance on tropes (the expressive ability of which is diminished by increased specificity). By examining the historical variations of such features, we may begin to reshape our understanding of periodization and its mechanisms.

Digital Lives: how people create, manipulate and store their personal digital archives

Peter Williams
University College London
peter.williams@ucl.ac.uk

Ian Rowlands
University College London
i.rowlands@ucl.ac.uk

Jeremy John
British Library
Jeremy.John@bl.uk

Introduction

The personal archives of scholars and other eminent people have long been kept in repositories such as the British Library, and include correspondence, notebooks, drafts of published papers, photographs etc. Needless to say, in recent years these collections have become ever more 'digital', as individuals are capturing and storing increasing amounts of digital information about or for themselves, including documents, portfolios of work, digital images, blogs, personal web pages and audio and video recordings (Summers & John 2001; Beagrie, 2005; John, 2005; Thomas & Martin, 2006). Not only the media and formats but also the *contents* of works created by individuals are changing. For example in the history of science, laboratory and field notebooks are changing drastically as research councils move towards supporting more standardised forms of recording e-science. We need to understand and address these issues now if future historians, biographers and curators are to be able to make sense of life in the early twenty-first century. There is a real danger otherwise that we will lose whole swathes of personal, family and cultural memory.

Personal information collections and management have been the focus of attention of researchers for many years, particularly in the field of Human-Computer Interaction (for an early example, see Malone, 1983). However, the rapid rise in digital applications and storage capacity has both stimulated interest in this field (see, e.g. Kaye et. al, 2006, Jones, 2007; Bruce et al, 2004) and also opened up a new research area within it directed at personal digital archives, a term used here to refer to these informal, diverse, and expanding memory collections created or acquired and accumulated and maintained by individuals and belonging to them, rather than to their institution or other place of work. It is the focus of a major study 'Digital Lives', which will be the subject of the presentation.

Digital Lives research project

The Digital Lives research project is being led by the British Library and funded by the UK's Arts and Humanities Research Council, focusing on such collections and their relationship with research repositories. It brings together expert curators and practitioners in digital preservation, digital manuscripts, literary collections, web-archiving, history of science, and oral history from the British Library with researchers in the School of Library, Archive and Information Studies at University College London and The Centre for Information Technology and Law at the University of Bristol. The project aims to explore how individuals manage their personal digital archives, in order to inform curators and archivists who will be entrusted with the personal collections of eminent people that will be left to them in future. We are seeking to clarify our understanding of an enormously complex and changing environment, engage with major issues, and evaluate radical new practices and tools that could assist curators in the future.

The initial phase of the project is complete, and will form the backdrop to the presentation. For this, we used in-depth interviews to explore the views, practices and experiences of a number of eminent individuals in the fields of politics, the arts and the sciences, plus an equal number of young or mid-career professional practitioners whose works and personal archives may be of interest to future scholars. Questions covered the history of the interviewee's experience with computers and ICT, training undertaken, manipulating files, backing up and transfer, collaborative work, extent and nature of any digital archive and the motivation behind decisions to keep or discard documents. Attitudes and perceptions of digital artifacts generally were also elicited.

This phase of the research (the preliminary findings of which have been reported by Williams *et al*, 2008) elicited a fascinating variety of experiences, behaviours and approaches, ranging from the poet who ignores word processing, electing instead to write all his drafts by hand, only committing the very final copy to a computer, to the geologist who forwards emails to himself so that the subject line can be changed, to the politician who has not digitised his audio diary but has it catalogued and cross-referenced online. Overall, the breadth of disciplines, backgrounds, ages and experiences of the individuals interviewed gave such contrasting and varied accounts that it is almost impossible to generalise findings. One thing is already very clear, however. The rise of Web 2.0 and document/data sharing and distribution over the internet has lead to large parts of individuals' digital assets (photographs, blog entries, even email archives) being hosted by servers geographically remote and with no guarantee of permanence by the service provider – creating immense problems for future curation and access. The amount of poorly named and inconsistently stored draft documents – sometimes with a 'final' version being lost in the digital ocean of one's 'C:' Drive will be another challenge.

Two surveys were undertaken, both online, developed from the issues raised and responses given in the qualitative phase of the research. One was to the general public, and the other to a sample of academics from both the humanities and sciences. We wished to explore differences in the online behaviours of different demographic and other groupings. It may be, for example, that scientists behave quite differently, in terms of their creation and management of their digital assets from humanities scholars; or that novice computer users operate in a fundamentally different way from experts. Initial analysis suggests, for example, that back-up and retrieval behaviour is different between genders. The presentation will integrate qualitative and quantitative findings from rigorous and sophisticated statistical analyses facilitated by SPSS software (quantitative) and from HyperResearch, (qualitative). This will provide an initial though deep and comprehensive examination of digital information behaviour.

Conclusion

It need only be said in conclusion, as the presentation will present findings from work presently in progress, that the way different people and groupings perceive, manipulate and store digital media has wide ranging implications for creative or scholarly activity, for the design and functionality of future software applications, and for the capture and care of personal digital archives by repositories and their future access by humanities scholars and others. The presentation should be of wide appeal to a 'digital humanities' audience.

References

Beagrie, N (2005), "Plenty of room at the bottom? Personal digital libraries and collections", *D-Lib Magazine,* 11(6) http://www.dlib.org/june05/beagrie.html

Bruce, H, Jones, W Dumais, S (2004) "Information behaviour that keeps found things found" *Information Research*, 10(1) paper 207 Available online: http://informationr.net/ir/10-1/paper207.html (accessed 12.10.08)

John, J L (2005) *Because topics often fade: letters, essays, notes, digital manuscripts and other unpublished works* pp399-422 in Narrow Roads of Gene Land Volume 3 Last Words, edited by M Ridley Oxford: Oxford University Press

Jones, W (2007) Personal information management In B Cronin (Ed.), *Annual review of information science and technology* Medford, NJ: Information Today pp453-504

Kaye, J, with J Vetis, A Avery, A Dafoe S David, L Onaga, I Rosero, T Pinch (2006) To have and to hold: exploring the personal archive *CHI Proceedings, Personal Information Management, April 22-27 2006* Montreal, 2006

Malone, T W (1983) how do people organize their desks? Implications for the design of office information systems *ACM Transactions on Office Information Systems* 1(1) pp99-112

Summers, A & John, J L (2001) The W D Hamilton Archive at the British Library *Ethology, Ecology & Evolution* 13, pp373-384

Thomas, S, & Martin, J (2006) Using the papers of contemporary British politicians as a testbed for the preservation of digital personal archives *Journal of the Society of Archivists*, 27(1) pp29-56

Williams P, Dean K, Rowlands I, John JL (2008) Digital Lives: report of interviews with the creators of personal digital collections *Ariadne* 55 http://www.ariadne.ac.uk/issue55/williams-et-al/ (posted 25.04.08; accessed 13.11.08)

Library as Agent of [Re]Contextualization

Vika Zafrin
Boston University
vzafrin@bu.edu

Jack Ammerman
Boston University
jwa@bu.edu

Garth Green
Boston University
gwgreen@bu.edu

Rapid production, dissemination and consumption of knowledge made possible in our digital culture is often accompanied by atomization of information. Once digitized, bits of information can easily be presented in a de-contextualized manner. This is both good and bad: any given datum lends itself more explicitly to being used in different, perhaps unexpected contexts; but without an originating context that is at least as widely known as the datum itself, it is difficult to find in the first place.

In his *Everything is Miscellaneous: The Power of the New Digital Disorder* (2007), David Weinberger has presented three orders – levels – of ordering. First order he likened to arranging books on a shelf, working with the objects themselves. Second order is creating a card catalog for the books – making objects to represent other objects, where each representative object refers to one and only one primary object (but the reverse doesn't hold: a book can have several catalog cards, filed by different criteria).

Third order does not presume an object to exist in only one place at a time. "The problems with the first two orders of order go back to the fact that they arrange atoms," Weinberger writes. "But now we have bits. [...] The third order removes the limitations we've assumed were inevitable in how we organize information."

This breakdown of "degrees of separation" between the things being ordered and the entities that order them is hardly new. In the 13th century, Bonaventure presented three possible "positions of thought": *extra se* (the mind views an outside object), *intra se* (a reflectio, or a 'thought about [an achieved] thought') and *supra se* (where we examine the principles behind our own classification and organization of observed knowledge). [*Itinerarium* 1, 4 and *Reductio* 10-12] Bonaventure merely systematized the tripartite structure of theology (sensible, symbolic and speculative) that he had inherited from such philosophers and theologians as Plotinus, Augustine, and pseudo-Dionysius.

The compelling common aspect of the third order of order, the *supra se* position of thought and the Absolute Idea, is recombination, re-iteration, a multiple contextualization of knowledge. Libraries have traditionally dwelt in the second order of order. At Boston University we have been playing with the notion of a library serving as a locus for third-order processes. This paper is a speculative discussion; we are re-conceiving library spaces as workshop-oriented, and collection development as a third-order activity based on the mash-up principle. We imagine the library as responsible for three spaces – physical, virtual and programmatic. This refashioned, multifaceted institution is an entity that not only enables research but engenders and even generates new knowledge by way of a new physical and conceptual structure.

Programmatic Space

Faculty members have expressed the desire for community in many places and contexts. What exactly do they mean, and is this something to which the library can respond? How do we create a physical community space for faculty in a large institution? How do we get people to engage with us and each other online?

Both the physical and the virtual spaces of the library, discussed below, are necessarily informed by a programmatic space – the conceptual framework within which we implement the plan to refashion the library. This includes both articulating goals (library as a space where new knowledge is generated) and experimenting with tools that may help us achieve those goals (user tagging for online catalog items). By imagining an environment that encourages direct interaction, and then suggesting starting discussion topics, we employ conversation theory (knowledge is created, and meanings agreed upon, in conversation) to refashion the library as a place that enables process as well as content.

Definitions and classifications have historically been created by field experts, who generally work within disciplinary boundaries, using disciplinary methods. Collaborative knowledge-gathering technologies allow definitions and structures to emerge from the materials themselves, potentially without regard to their prior classification(s). This emergent knowledge, created within programmatic space, informs the way in which reconceptualizing the virtual space will translate into transforming the library's physical space.

Physical / Virtual Space

We put practical considerations such as funding aside for the moment: before proposing a capital renovation project to the university, we must have a clear idea of what we are trying to accomplish.

Most library users associate the library's physical space with silent study. With clever layout planning — perhaps re-shuffling stacks to de-emphasize them and minimize sound movement — a library can accommodate multiple spaces for quiet group discussion in addition to carrels. Round tables are provided, with computers and large-screen monitors for collaborative viewing and discussion of online resources. The library becomes a constant low-grade workshop space that makes both physical and virtual materials equitably available.

We are currently creating two topical online resources, which follow from our relatively new understanding of the library itself as a pedagogical tool. Our History of Missiology site [1] aims to present an extensive and unique collection of letters, journals, biographies, travel descriptions, and books written by and about protestant missionaries. These materials will be of use both to theologians and to social scientists, as they constitute the earliest anthropological records of several cultures.

We are also gathering resources about the Personalism school of thought, which flourished at BU in the early 20th century. It has played an important role in the development of philosophical theology, but its implications are far-reaching in their interdisciplinarity. Personalism has not been adequately studied, and connections between it and later philosophical thought have not been explicitly drawn. By providing primary materials and tools for discussion, we again hope to facilitate emergent knowledge.

Collection Development and Cataloging

Approaching collection development as a third-order activity, we consider the number of books on shelves one, but not the sole, determining factor of the strength of a collection. Programmatic restructuring of the library's physical and virtual spaces is aimed at gathering interdisciplinary groups, which in turn will ideally participate in collection development along with library staff. As knowledge can be built up, mapped, by free (informed) association, so library collections can be developed in a similar way.

Thus, a series of meetings dedicated to the revival of Personalism may attract theologians, philosophers, historians and sociologists. Each scholar will bring knowledge of different relevant resources to the table, and in the course of discussion draw out a coherent and comprehensive body of knowledge that the community believes to be essential to its work. This is new in that we aim to gather people in the library, both as a physical space and as a concept within the larger structure of a research institution, in a purposeful and sustained way. Through this process we would create a more participatory and connected relationship both between researchers across their disciplinary boundaries, and between the library and their respective research programs.

In the past few decades, with the advent of electronic cataloging, libraries have been slowly moving towards the third order of order. For the most part libraries still catalog physical objects; it is only very recently that we have been able to begin digitizing those objects, and have had to face the existence of born-digital artifacts also in need of cataloging. Momentum is difficult to influence: even with digital objects, libraries fall into patterns of cataloging typical of physical artifacts. Developing referral systems, social tagging, etc. will be required for an accurate representation of these complex and structurally novel entities.

Moving Forward: Assessment and Implementation

In order for the project of refashioning the research process to succeed, a crucial component is an outreach coordinator who is also a scholar in the field. This would be a person who rallies the researcher troops and continually spurs conversation. Without such a constant external stimulus, and with so many other demands on their time and attention, researchers are less likely to participate in – or initiate – discussion.

This sort of rallying can and should be institutionalized. With a dedicated driving force behind them, brown-bag seminars, lecture/presentation series and discussion lists can be a thriving ground for exploring ideas. Such efforts have so far been led by dedicated individuals (for example, Willard McCarty in his stewardship of the Humanist discussion list) and often at digital humanities centers (the MITH Digital Dialogues, the Computers in the Humanities Users Group at Brown, the Seminar in Humanities Computing at the Centre for Computing in the Humanities in King's College London, and many others). The library as an institution is well positioned to join in as a venue for scholarly communication, given adequate staffing and support.

How to actually accomplish all of these things? We will present the challenges we have faced thus far, and our approach to overcoming those challenges during the current academic year, by addressing several recommendations from the final report of the ACLS Commission on Cyberinfrastructure, released in 2006.

[1] http://digilib.bu.edu/mission/

The Artificial Intelligence (AI) Hermeneutic Network: Toward an Approach to Analysis and Design of Intentional Systems

Jichen Zhu
Georgia Institute of Technology
jichen.zhu@lcc.gatech.edu

D. Fox Harrell, Ph.D.
Georgia Institute of Technology
fox.harrell@lcc.gatech.edu

> 'I felt that I should be able to get the computer to sound good more or less on its own, so that someone listening to it says, "Who is that playing?" But if you get "What's that?" instead, you have to go back to the drawing board.' (Lewis, 2000)

Abstract

Digital information technologies are increasingly being adopted in the humanities as both research tools and supports for new forms of cultural expression. Some of these digital technologies, in particular artificial intelligence (AI) programs, exhibit complex behaviors usually seen as the territory of intentional human phenomena, such as creativity, planning and learning. This paper identifies a prototypical subset of these programs, which we name *intentional systems*, and argues that their seemingly *intentional* behaviors are not the sole effect of underlying algorithmic complexity and knowledge engineering practices from computer science. In contrast, we argue (paralleling the field of software studies) that intentional systems, and digital systems at large, need to be analyzed as a contemporary form of historically, culturally, socially, and technically situated *texts*. Perception of system intentionality arises from a network of continuous meaning exchange between system authors' narration and users' interpretation processes embedded in a broader social context. The central contribution of this paper is a new interdisciplinary analytical framework called the *AI hermeneutic network* that is informed by traditions of hermeneutic analysis, actor-network theory, cognitive semantics theory, and philosophy of mind. To illustrate the design implication of the AI hermeneutic network, we present our recent work *Memory, Reverie Machine*, an expressive intentional system that generates interactive narratives rich with daydreaming sequences.

Intentional Systems

Trombonist and composer George Lewis's above description of his interactive musical system *Voyager* exemplifies a growing number of digital systems, such as the autonomous painting program *AARON* (Cohen 2002) and recent computational narrative works (Harrell 2006; Mateas & Stern 2002; PŽrez y PŽrez & Aliseda 2006), that utilize AI techniques in pursuit of cultural expression. Decades after heated debates about the feasibility of AI, the question of whether computers may one day possess human-level intelligence no longer spurs society's fear and curiosity. Instead, systems are designed to encourage users to make sense of them as intentional and independent entities. Compared to instrumental, production-oriented programs such as the *PhotoShop*, these systems display intentional behaviors related to human mental phenomena such as planning, learning, narrating, and creating, as if their actions were *about* something in the world (Searle 1983) rather than mere execution of algorithmic rules. Lewis, for instance, insists that *Voyager* 'not [be] treated as a musical instrument, but as an independent improviser.' He deliberately designed the system to display independent behaviors arising from its own internal processes that even its designer cannot fully anticipate. The improvisational *dialogue* between *Voyager* and the musicians, Lewis emphasizes, is 'bi-directional transfer of intentionality through sound.' computational complexity, 2) process opacity, 3) human-like coherent behaviors, and 4) execution of authorial intention. The term encompasses not only AI systems but also AI-like systems that exist either outside of computer science communities or are not described by their authors as AI systems for ideological or other reasons. Critical analysis and design of intentional systems, like information technologies at large in the digital humanities, calls for the recognition of these systems as important forms of cultural production, beyond their traditionally instrumentalized, productivity oriented roles.

Intentional Systems as Texts

Although generally used to describe written forms of discourse, the term *text* as the object of literary theory and modern hermeneutics is not confined to only linguistic forms. In his essay on the literary text, German philosopher Manfred Frank (Frank 1989) criticizes the notion that meanings that authors encode within texts can be objectively retrieved without distortion by readers given appropriate methods of interpretation (Hirsch 1967). Instead, Frank proposes a complex communication process in which both author and reader actively create, shape, and reconstruct meanings. This echoes the even broader notion of dialogic meaning posited by the Russian philosopher and critic Mikhail Bakhtin in which language is understood as dynamic, contextual, intertextual, and relational (Holquist 1990). Acknowledging the textuality of intentional systems opens up understanding of system intentionality to a range of socially situated methods.

Intentional systems are not simply the result of clever algorithmic and data structural innovations. The AI practitioner and theorist Philip Agre cogently points out that the 'the purpose of AI is to build computer systems whose operation can be narrated using intentional vocabulary.' (Agre 1997) Michael Mateas, co-developer of *Façade,* further deconstructs the codes invoked in AI practice by computation, and definitions of system progress) and the co-existing 'code machine' (including physical processes, computational processes, and complex causal flow), in order to pin down the long-neglected social and discursive aspect of AI systems (Mateas 2002). In addition to considering actual computer programs, analysis of intentional systems should not omit the authors' publications, presentations, and interpersonal communication about the system. Such narrative outputs situate the system in AI research communities and frame users' interpretation, and therefore must be considered as part of the intentional system.

The AI Hermeneutic Network

The central contribution of this paper is the *AI hermeneutic network* model, enabled by theorizing intentional systems as texts. The interdisciplinary framework analyzes system intentionality as a result of a hermeneutic communication process that involves both authors' narrations and users' interpretations through interaction with both actual systems and authors' narrative output. In addition, this paper recognizes that intentional systems exist in broader social contexts that involve more than just authors and users. Animate and inanimate actors, called 'actants' in actor-network theory (Callon 1986; Latour 1996), participate in the network through multi-directional communication. Government and military funding, for instance, often plays a prominent role in determining direction and validity of different approaches of AI research.

Historically, hermeneutic studies developed interpretative theories and methods in order to recover the meanings of sacred texts intended by the (divine) author(s). Modern hermeneutics, influenced by Schleiermacher, recognizes that *everything* calls for the work of interpretation and broadens itself to the philosophical interrogation of interpretation (Virkler 1981). This paper highlights discursive 'elasticity' of the AI key words, such as planning (Agre 1997). He observes that these key terminologies are simultaneously precise (formal) and vague (vernacular), which allows AI practitioners to seamlessly integrate their everyday experience as embodied intentional being in the algorithmic research, and to narrate computation with popularly accessible vernacular vocabulary.

One relatively unexplored aspect of this continuous negotiation of values and meanings between both human and computational actors (Latour 1996) is users' readings and interpretations of intentionality from systems that are clearly inanimate. For example, human co-performers and their audiences' interpretations of *Voyager*'s behaviors as intentional are central to construe the systems' status as an independent performer in its own right, as intended by its designer. Frank argues that '[i] n the understanding of its readers the text … acquires a meaning which exceeds the memory of its origin.'(Frank 1989) Any analysis of system intentionality then is not adequate without considering participation of users and audiences.

This paper emphasizes the discursive strategy and semantic interpretation from a cognitive linguistics perspective. Conceptual blending theory (Fauconnier 2001; Fauconnier & Turner 2002; Turner 1996) offers a cognitive foundation for understanding system intentionality as actively (re)constructed by users via integrating concepts of intentionality based on encounters with animate agents, and conceptualization of algorithmic operation of inanimate computer systems. Thus, users compress the behavior of unfamiliar computational systems to human scale by constructing conceptual blends of systems with human-like intentionality, through semantic hooks that facilitate such blends in the various discourses surrounding the systems.

Conclusion: Design Implications of the AI Hermeneutic Network

The novel framework of the hermeneutic network suggests new design approaches for intentional systems in digital humanities. Our current interactive narrative work *Memory, Reverie Machine* generates stories in which the main character varies dynamically along a scale between a user-controlled avatar with low intentionality and an autonomous non-player character with high intentionality. By algorithmically controlling the semantic hooks for interpreting system behavior as intentional in the narrative discourse (Zhu & Harrell 2008), the authors turn system intentionality into a scalable expressive dimension in interactive storytelling (Harrell & Zhu 2009).

In conclusion, this paper proposes a new interdisciplinary framework to analyze intentional systems as social and cultural productions, as opposed to construing them as the domain of purely technical practices. It underlines authors' narrative and users' interpretative strategies, in a socially situated network of meaning exchange. Finally, through our own computational work we suggest new design implications for intentional systems, such as the scale of intentionality (Zhu & Harrell 2008) that potentially can add new forms of expressivity to intentional systems in digital humanities.

References

Agre, P. E. (1997). 'Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI', in *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*, eds G. C. Bowker, S. L. Star, W. Turner, L. Gasser & G. Bowker, Lawrence Erlbaum Associates, pp. 131-58.

Callon, M. (1986). 'Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay', in *Power, Action and Belief: A New Sociology of Knowledge*, ed. J. Law, Routledge & Kegan Paul, London.

Cohen, H. (2002). 'A Self-Defining Game for One Player: On the Nature of Creativity and the Possibility of Creative Computer Programs', *Leonardo*, vol. 35, no. 1, pp. 59-64.

Fauconnier, G. (2001). 'Conceptual Blending and Analogy', in *The Analogical Mind: Perspectives from Cognitive Science*, eds D. Gentner, K. J. Holyoak & B. N. Kokino, MIT Press, Cambridge, MA.

Frank, M. (1989). *The Subject and the Text: Essays on literary theory and philosophy*, Cambridge University Press, Cambridge.

Harrell, D. F. (2006). Walking Blues Changes Undersea: Imaginative Narrative in Interactive Poetry Generation with the GRIOT System, paper presented to AAAI 2006 Workshop in Computational Aesthetics: Artificial Intelligence Approaches to Happiness and Beauty, Boston, MA, AAAI Press.

Hirsch, E. D. (1967). *Validity in Interpretation* Yale University Press, New Haven and London.

Holquist, M. (1990). *Dialogism: Bakhtin and His World*, 2 edn.

Latour, B. (1996). *Aramis, or the Love of Technology*, Harvard University Press, Cambridge.

Mateas, M. (2002). Interactive Drama, Art, and Artificial Intelligence, Carnegie Mellon University.

Searle, J. (1983). *Intentionality: An Essay in the Philos-*

ophy of Mind, Cambridge University Press, Cambridge.

Turner, M. (1996). *The Literary Mind: The Origins of Thought and Language*, Oxford UP, New York; Oxford.

Virkler, H. A. (1981). *Hermeneutics: Principles and Processes of Biblical Interpretation*, Baker Book House, Grand Rapids, MI.

Fauconnier, G. & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*, Basic Books.

Harrell, D. F. & Zhu, J. (2009). Agency Play: Dimensions of Agency for Interactive Narrative Design, paper presented to to appear in the Proceeding of AAAI Spring 2009 Symposium on Interactive Narrative Technologies II.

Mateas, M. & Stern, A. (2002). 'A Behavior Language for Story-Based Believable Agents', *IEEE Intelligent Systems*, vol. 17, no. 4, pp. 39-47.

Pérez y Pérez, R. & Aliseda, A. (2006). The Role of Abduction in Automatic Storytelling, paper presented to Proceedings of the AAAI workshop in Computational Aesthetics: AI Approaches to Beauty & Happiness, Boston, MA, AAAI Press, pp. 53-60.

Zhu, J. & Harrell, D. F. (2008). Daydreaming with Intention: Scalable Blending-Based Imagining and Agency in Generative Interactive Narrative, paper presented to AAAI 2008 Spring Symposium on Creative Intelligent Systems, Stanford, CA, AAAI Press, pp. 156-62.

Posters

SEASR integrates with Zotero to Provide Analytical Environment for Mashing up Other Analytical Tools

Loretta Auvil
University of Illinois at Urbana-Champaign
lauvil@ncsa.uiuc.edu

Boris Capitanu
University of Illinois at Urbana-Champaign
capitanu@ncsa.uiuc.edu

Xavier Llorà
University of Illinois at Urbana-Champaign
xllora@illigal.ge.uiuc.edu

Michael Welge
University of Illinois at Urbana-Champaign
welge@ncsa.uiuc.edu

Bernie Ács
University of Illinois at Urbana-Champaign
bernie@ncsa.uiuc.edu

This paper describes a development effort to link two humanities cyberscholarship infrastructure projects supported by The Andrew W. Mellon Foundation. We have created an extension to Zotero [1] that acts as a bridge between the data stored by Zotero, and the suite of analytic tools provided by SEASR [2]. This extension provides users with the ability to apply a variety of data analysis algorithms to their Zotero constructed collections, and visualize the results directly in the browser. This is accomplished by directly accessing the data model provided by Zotero, and converting that data model into RDF, which allows the ability to exploit the analytical capabilities of SEASR.

The SEASR environment provides a framework to integrate data, analytics, and tool constructs, so that data from one component can be passed to another. One of the unique capabilities of SEASR is the facility to provide a tool for mashups. That is, the ability to allow users to combine tools in efficient and effective ways. This paper describes the coupling of two relevant environments for humanist, Zotero and SEASR - Zotero's data asset creation with the analytical capabilities of SEASR. Through the use of Zotero's plugin environment, we can execute the analysis capabilities of the SEASR environment.

The following sections provide a description of the two major pieces of this effort—Zotero and SEASR. Also provided is a description of the major functions performed by the combination of the two. These include: data gathering, data analytics, and data visualization. We end with a summary of the integration of the two efforts and a view to our future work.

1. Background

1.1 Zotero

Zotero was selected because of its popularity with scholars to record, catalog and find resources collected from the Internet. Zotero was developed at the Center for History and New Media, George Mason University, and is a tool aimed at facilitating a user's research process by providing mechanisms for collecting, managing, and citing internet resources. Zotero functions as an extension of the popular open-source browser, Firefox, which allows it to provide its services in the same environment where the research is usually performed. One of the key features provided by Zotero is the ability to automatically extract metadata from online resources as part of the resource collection process, and store it conveniently on the user's computer, allowing for offline retrieval of this data on demand. Zotero also provides advanced tagging and searching functionality, allowing the user to organize, find, and visualize the collected resources.

Zotero includes a powerful metadata editor, allowing the user to make additions/corrections to the automatically extracted information. Users can add new fields, attach screenshots and documents, create notes, and even create relationships between the various resources collected.

Overall, with such a vast and diverse amount of information, a mechanism for finding patterns or interesting relationship between these resources would go a step further in helping researchers discover and extract more information from their collections. Enter SEASR.

1.2 SEASR
(Software Environment for the
Advancement of Scholarly Research)

SEASR analytics enhances scholars' use of digital materials by helping them uncover hidden information and connections, supporting the study of assets from small patterns drawn from a single text or chunk of text to broader entity categories and relations across a million words or a million books. SEASR is designed to enable digital humanities scholars to rapidly design, build, and share software applications that support research and collaboration.

The SEASR team developed Meandre [3], which is the machinery for assembling and executing data flows—software applications consisting of software components that process data (such as by accessing a data store, transforming the data from that store and analyzing or visualizing the transformed results).

SEASR is extensible allowing for new analytics to be added, such as support for linguistic analysis for different time periods or languages, to readjusting entire steps in the work process so that researchers can validate results from their queries. Components can be created from other programming projects. The SEASR environment is data driven and includes a workbench to orchestrate the flow of data through the different components. All SEASR analytics are enabled as web service calls.

SEASR also provides publishing capabilities for flows and components, enabling users to assemble a repository of components for reuse and sharing. This allows users to leverage other research and development efforts by querying and integrating component descriptions that have been published previously at other shareable repository locations.

2. Data Gathering

Zotero's data model is very flexible, allowing the user to add new fields, create notes, attach documents and screenshots, and establish relationships between resources. At a minimum, Zotero adds the following information for each resource that is added: title of the resource, originating URL, and the dates when the resource was created, modified, and accessed. For many major research and library sites Zotero can automatically extract the full reference information, which includes authorship data, abstracts, page references, locations, etc. This provides a wealth of information that can then be submitted for analysis to SEASR.

Once the data are converted into RDF (a process which is transparent to the user), it can be sent for processing, at user's request, to any of a number of available data analysis algorithms. When such a data processing request is received, the extension establishes a communication channel with the web service associated with the processing flow, through which the RDF data are submitted. After processing completes, the results are retrieved via the same communication channel and displayed in a new browser window. Depending on the complexity of the type of processing requested, there may be a significant delay until the results are retrieved.

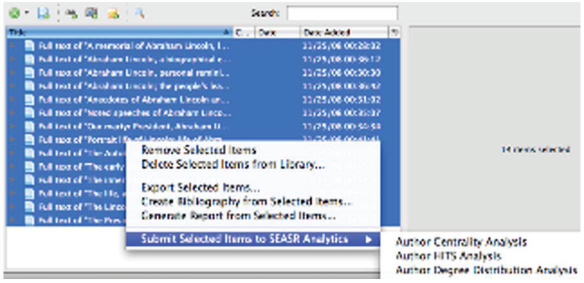

Search:
Title
Date
Date Added
Remove Selected Items
Delete Selected Items from Library...
Export Selected Items...
Create Bibliography from Selected Items...
Generate Report from Selected Items...
Submit Selected Items to SEASR Analytics
Author Centrality Analysis
Author HITS Analysis
Author Degree Distribution Analysis

Figure 1. Plugin for Zotero that requests SEASR analytics.

The extension provides a flexible mechanism through which the user can specify which data processing flows they want to have access to, by configuring a list of SEASR servers where these flows are hosted. This way, the user can include any number of Zotero-compatible data processing flows hosted by 3rd party organizations.

3. Data Analytics

The SEASR team has been integrating a variety of tools as well as developing our own analytics. Currently we have integrated natural language processing tools (NLP) and current research algorithms from our data mining collaborators as well as transformation components to allow for data movement between the different components.

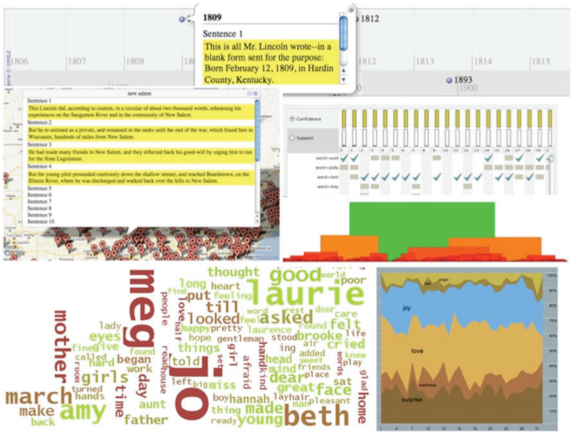
We have enabled some very simple and straightforward requests, like word counts, information regarding part of speech, and entity extraction capabilities. We also have additional machine learning approaches that can be leveraged, like clustering, frequent pattern analysis, predictive modeling, graph mining, and sequence analysis. We have currently integrated D2K (Data to Knowledge) [4] and T2K (Text to Knowledge) analysis, OpenNLP [5], and GATE (General Architecture for Text Engineering) [6]. This means that from your Zotero collection, you can ask for a social network analysis based on authors and other metadata. You can ask for a tag cloud of all your notes. You can ask for a tag cloud of a particular work or collection. You can cluster the documents from your collection. You can track a character or set of terms throughout a book or collection. You can look at extracted entities like locations on a Google map [7]. You can look at extracted entities like date on a timeline like Simile [8]. You can build a social network of the people mentioned in your collection.

4. Data Visualization

As with the data analysis, a number of visualization tools exist, so we have been working to integrate with these tools rather than redeveloping. We have incorporated visualizations from D2K as applets such as frequent pattern analysis as well as several of the predictive modeling visualizations. We have also leveraged code to create a tag cloud [9]. We are providing link-node charts and stacked bar charts via flare [10]. The collage of example visualizations below is meant to provide an idea of the visual metaphors being used.

5. Future Work

We continue developing analysis and visualization capabilities that can be leveraged by Zotero. As part of our Pathways to SEASR Workshops, we are demonstrating this tool integration and are establishing collaborations with workshop teams. These teams are exploring specific use cases to demonstrate scholarly research that can be easily added to this plugin environment. We are looking to improve the interaction between the plugin and the SEASR framework and its ability to provide users with visual interfaces for customizing the execution of flows.



1809

Sentence 1

This is all Mr. Lincoln wrote--in a blank form sent for the purpose: Born February 12, 1809, in Hardin County, Kentucky.

new salem

Sentence 1

This Lincoln did, according to custom, in a circular of about two thousand words, rehearsing his experiences on the Sangamon River and in the community of New Salem.

Sentence 2

But he re-enlisted as a private, and remained in the ranks until the end of the war, which found him in Wisconsin, hundreds of miles from New Salem.

Sentence 3

He had made many friends in New Salem, and they reflected back his good-will by urging him to run for the State Legislature.

Sentence 4

But the young pilot proceeded cautiously down the shallow stream, and reached Beardstown, on the Illinois River, where he was discharged and walked back over the hills to New Salem.

Sentence 5

Sentence 6

Sentence 7

Sentence 8

Sentence 9

Sentence 10

Figure 2. Collage of visual metaphors available with SEASR analytics.

6. Summary

In summary, we have created a tool that facilitates the communication of Zotero collections data with SEASR for further study and research. We have linked SEASR, a strong and flexible tool that can add research capabilities to these text assets. SEASR allows for the use of its existing analysis and visualization tools, and more, it allows for the integration of other tools through a mashup process. The result of this effort is a synergy—a strengthening of both Zotero and SEASR—as useable tools for cyberscholarship.

7. Acknowledgement

The SEASR project is funded by The Andrew W. Mellon Foundation.

8. References

1. Zotero, http://www.zotero.org

2. SEASR, http://seasr.org

3. Llorà, Ács B, Auvil LS, Capitanu B, Welge ME, Goldberg DE (2008) Meandre: Semantic-Driven Data-Intensive Flows in the Clouds, in Proceedings of IEEE Fourth International Conference on eScience, 238-245, IEEE Press.

4. D2K, http://alg.ncsa.uiuc.edu/do/tools/d2k

5. OpenNLP, http://opennlp.sourceforge.net/

6. H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

7. Google Maps API, shttp://code.google.com/apis/maps/

8. Simile Timeline, http://www.simile-widgets.org/timeline/

9. Tag Cloud, http://emumarketing.uoregon.edu/paul/2008/09/28/the-new-tag-cloud/

10. Flare, http://flare.prefuse.org/

Automatic Standardization of Spelling for Historical Text Mining

Alistair Baron
Lancaster University
a.baron@comp.lancs.ac.uk

Paul Rayson
Lancaster University
p.rayson@lancs.ac.uk

Dawn Archer
University of Central Lancashire
dearcher@uclan.ac.uk

Introduction

The use of textual data in humanities research is significantly aided by automated techniques such as key word analysis, collocations and corpus annotation (e.g. part-of-speech). If a text corpus contains a large amount of spelling variation, there is a considerable impact on the accuracy of these automatic techniques. For example, studies in respect to Early Modern English (EModE) corpora - the focus of the study detailed in this paper - have documented the adverse effects of spelling variation on key word analysis (Baron et al., 2009), part-of-speech tagging (Rayson et al., 2007) and semantic analysis (Archer et al., 2003).

The problem of spelling variation in corpora needs to be addressed in order for more accurate and meaningful results to be achieved in fields where historical source texts are required. Researchers can side-step the issue by using modernized versions of corpora, of course, but these are not always available. Another potential solution is to manually standardize the spellings; this includes reading through texts, spotting any non-standard spellings and deciding upon a modern equivalent, resulting in the production of a new version of the text with spelling variants replaced. However, a manual standardizing approach is likely to be unworkable when working with some of the larger corpora or online databases that are now available.

This paper details the current version of the VARiant Detector (VARD 2) tool, which can be used in various ways to standardize spelling variation in corpora. In particular, the tool can be used to (partially) standardize spellings automatically, with no restriction on the number of words to be processed. Here, we focus on the ways in which the tool can be trained from manually standardized corpora samples, particularly the letter replacement component of the tool, and evaluate the improvement that this makes to the performance of VARD 2.

Early Modern English Spelling Variation

The EModE period is of particular interest in historical text mining studies; book production increased sharply during the period, largely due to the introduction of the printing press (1476) and increasing literacy levels (Görlach, 1991: 6). As such, the EModE period is the earliest period of the English Language from which a reasonably large corpus can be constructed and studied in detail.

Spelling variation was a major feature of EModE texts, the extent of which we have recently quantified in Baron et al. (2009). It is common to find words spelt in a number of different forms in the same text or even on the same page. This was not seen as problematic, however, as there was no notion of the importance for a single spelling for each word; for example, letters would be added or removed to ease line justification. Table 1 below shows some typical spelling variants found in EModE texts, whilst Vallins and Scragg (1965) and Culpeper and Archer (forthcoming) describe the spelling variation trends in more detail.

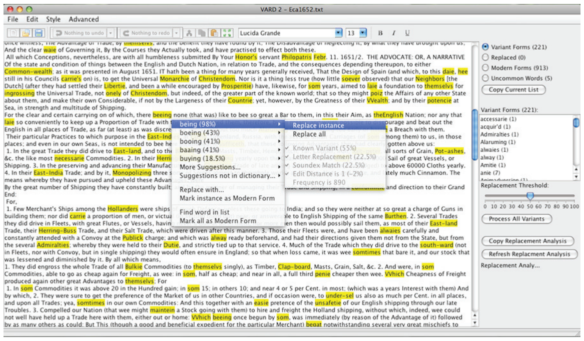
Variant	Modern Equivalent	Notes
"goodnesse"	"goodness"	'e' often added to end of words.
"brush'd"	"brushed"	Apostrophes used instead of 'e'.
"encrease"	"increase"	Vowels often interchanged.
"spels"	"spells"	Consonants often doubled or singled.
"deliuering"	"delivering"	Common for 'u' and 'v' to be interchangeable.
"conuay'd"	"conveyed"	Many combinations of the above.

Table 1. Examples of common EModE spelling variants.

We have shown the effect of this spelling variation on textual analysis techniques in previous and forthcoming papers: key word analysis (Baron et al, 2009), part-of-speech tagging (Rayson et al., 2007) and semantic tagging (Archer et al., 2003). All of the studies showed that spelling variation causes considerable problems to the accuracy and meaningfulness of results, and that dealing with spelling variation (even partially) can achieve substantial improvements in annotation accuracy.[1] The production of standardized or modernized versions of historical corpora therefore allows for more accurate automated text mining techniques to be applied to the corpora.[2]

VARD 2 and DICER

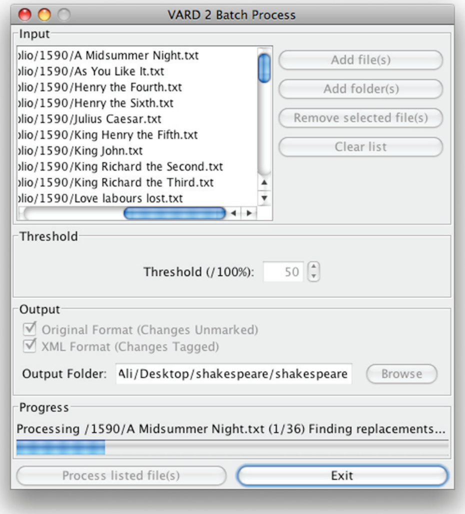
Our solution to the spelling variation problem described in the previous section has been the development of the VARD 2 tool,[3] a piece of software designed to assist researchers in standardizing historical corpora (specifically EModE texts) both manually and automatically. VARD 2 uses a manually created list of variant – replacements mappings as well as employing methods from modern spell checking software; such as phonetic matching, letter replacement heuristics and Edit Distance. An earlier version of the VARD 2 software is described in more detail and evaluated in Rayson et al (2008). The current version can cater for user-created letter replacement rules, which will be used by the tool to find potential variant replacements. In addition, XML provision has been improved, processing speed increased, and a new word reference list[4] added. Screenshots of the latest version, VARD 2.2, are shown in Fig. 1 and Fig. 2.



once witness, The Advantage of Trade, by themselvs, and the benefit they have found by it, The Disadvantage of neglecting it, by what they have brought upon us,
And the clear waie of Governing it, By the Courses they Actually took, and have practised to effect both these.
All which Conceptions, nevertheless, are with all humbleness submitted By Your Honor's servant Philopatris Febr. 11. 1651/2. THE ADVOCATE: OR, A NARRATIVE
Of the state and condition of things between the English and Dutch Nation, in relation to Trade, and the consequences depending thereupon, to either
Common-wealth; as it was presented in August 1651. IT hath been a thing for many years generally received, That the Design of Spain (and which, to this daie, hee
still in his Councils carrie's on) is, to get the Universal Monarchie of Christendom. Nor is it a thing less true (how little soever observed) that our Neighbors [the
Dutch] (after they had settled their Libertie, and been a while encouraged by Prosperitie) have, likewise, for som years, aimed to laie a foundation to themselvs for
ingrossing the Universal Trade, not onely of Christendom, but indeed, of the greater part of the known world; that so they might poiz the Affairs of any other State
about them, and make their own Considerable, if not by the Largeness of their Countrie; yet, however, by the Greatness of their VVealth; and by their potencie at
Sea, in strength and multitude of Shipping.
For the clear and certain carrying on of which, there beeing none (that was) like to bee so great a Bar to them, in this their Aim, as theEnglish Nation; nor any that
laie so conveniently to keep up a Proportion of Trade with ... ourage and beat out the
English in all places of Trade, as far (at least) as was discre ... a Breach with them.
Their particular Practices to which purpose in the East-Ind ... mong them) to us, in those
places; and even in our own Seas, is not intended to bee he ... gotten above us:
1. In the great Trade they did drive to East-land, and to th ... ll sorts of Grain, Pot-ashes,
&c. the like most necessarie Commodities. 2. In their Herri ... Sail of great Vessels, or
Shipping. 3. In the preserving and advancing their Manufac ... above 60000 Cloths yearly.
4. In their East-India Trade; and by it, Monopolizing three ... ately much Cinnamon. The
means whereby they have pursued and upheld these Advan
By the great number of Shipping they have constantly built ... and direction to their Grand
End:
For,
1. Few Merchant's Ships among the Hollanders were ships ... India; and so they were neither at so great a charge of Guns in
building them; nor did carrie a proportion of men, or victua ... le to English Shipping of the same Burthen. 2. Several Trades
they did drive in Fleets, with great Flutes, or Vessels, havin ... en then would possibly sail them, as most of their East-land
Trade, their Herring-Buss Trade, and their Salt Trade, which were driven after this manner. 3. Those their Fleets were, and have been alwaies carefully and
constantly attended with a Convoy at the Publick charge; and which was alway ready beforehand, and had their directions given them not from the State, but from
the several Admiralties; whereby they were held to their Dutie, and strictly tied up to that service. 4. Much of the Trade which they did drive to the south-ward (not
in Fleets, nor with Convoy, but in single shipping) they would often ensure in England; so that when loss came, it was wee somtimes that bare it, and our stock that
was lessened and diminished by it. By all which means,
1. They did engross the whole Trade of all Bulkie Commodities (to themselvs singly), as Timber, Clap-board, Masts, Grain, Salt, &c. 2. And were, in som
Commodities, able to go as cheap again for Freight, as wee: in som, half as cheap; and near in all, a full third penie cheaper then wee. VVhich Cheapness of Freight
produced again other great Advantages to themselvs; For
1. In som Commodities it was above 20 in the Hundred gain; in som 15; in others 10; and near 4 or 5 per Cent. in most: (which was a years Interest with them) And
by which, 2. They were sure to get the preference of the Market of us in other Countries, and if occasion were, to under-sel us also as much per Cent. in all places,
and upon all Trades; yea, somtimes in our own Commodities: And this together with an easie pretence of the unsafetie of our English shipping through our late
Troubles. 3. Compelled our Nation (that wee might maintein a Stock going with them) to hire and freight the Holland shipping, without which, indeed, wee could
not well have held up a Trade here with them, either out or home: VVhich beeing once begun by som, was immediately (by reason of the Advantage of it) followed
by as many others as could: But This (though a good and beneficial expedient for the particular Merchant) begat notwithstanding several very great mischiefs to

Fig. 1 Screenshot of VARD 2.2 showing the interactive mode which allows the user to manually standardize texts and train the tool on samples of a corpus

One way in which VARD 2 can be used is to automatically standardize the spelling variation in an entire corpus. For EModE texts, this can be done immediately, with no training. However, for better results and to use the tool with other varieties of English, the user can train the software on a particular corpus by using the interactive version to manually process samples from the corpus. The tool will improve its ability to deal with a corpus based on decisions made by the user in the interactive version. It does this by learning which of its methods are most successful in finding the correct replacement for variants and adjusting its method weights accordingly (these are used when ranking potential replacements). The tool will also edit its dictionary and its list of specific variant replacements based on changes made by the user.



plio/1590/A Midsummer Night.txt
plio/1590/As You Like It.txt
plio/1590/Henry the Fourth.txt
plio/1590/Henry the Sixth.txt
plio/1590/Julius Caesar.txt
plio/1590/King Henry the Fifth.txt
plio/1590/King John.txt
plio/1590/King Richard the Second.txt
plio/1590/King Richard the Third.txt
plio/1590/Love labours lost.txt

Threshold (/100%): 50

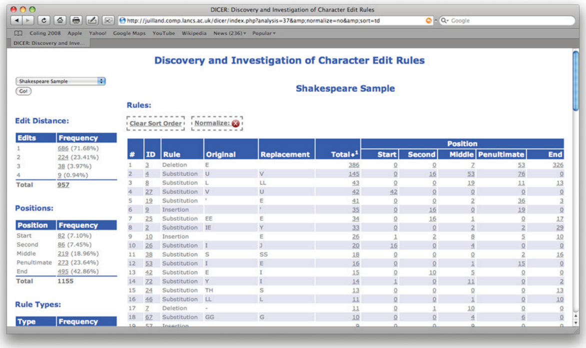
☑ Original Format (Changes Unmarked)
☑ XML Format (Changes Tagged)

Output Folder: Ali/Desktop/shakespeare/shakespeare

Processing /1590/A Midsummer Night.txt (1/36) Finding replacements...

Fig. 2 Screenshot of VARD 2.2 showing the batch-processing mode where users can automatically standardize a chosen group of texts

A new development to allow for further training of VARD 2 on a corpus is a tool named DICER (Discover and Investigation of Character Edit Rules). DICER can search XML output from VARD 2 for variant – replacement mappings or be provided with a list of such mappings. Each mapping is analyzed and a set of character edit rules are produced which can transform the spelling variant into its modern equivalent. The details of these character edit rules are then collated into a database, which can be viewed through a set of web pages.[5] The main table produced by the analysis, shown in Fig. 3, displays details of the individual character edit rules along with various frequencies. By clicking on individual rules, further information is available such as which characters typically occur before and after the rule occurs; this is shown for the rule 'Delete e' in Fig. 4. Any frequency in the tables can be clicked to view a list of occurrences producing that frequency. The data produced in the DICER analysis is vast and thus cannot be detailed in full here.


DICER: Discovery and Investigation of Character Edit Rules
http://juilland.comp.lancs.ac.uk/dicer/index.php?analysis=37&amp;normalize=no&amp;sort=td
Google
Coling 2008
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News (236)
Popular
DICER: Discovery and Inve...
Shakespeare Sample
Go!
Clear Sort Order
Normalize:

Discovery and Investigation of Character Edit Rules

Shakespeare Sample

Edit Distance:

Edits	Frequency
1	686 (71.68%)
2	224 (23.41%)
3	38 (3.97%)
4	9 (0.94%)
Total	957

Positions:

Position	Frequency
Start	82 (7.10%)
Second	86 (7.45%)
Middle	219 (18.96%)
Penultimate	273 (23.64%)
End	495 (42.86%)
Total	1155

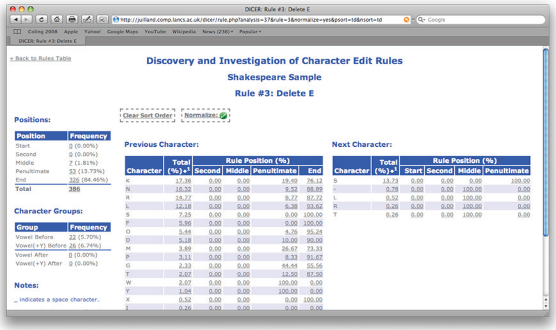
Rule Types:

Type	Frequency

Rules:

#	ID	Rule	Original	Replacement	Total↓1	Position				
						Start	Second	Middle	Penultimate	End
1	3	Deletion	E		386	0	0	7	53	326
2	4	Substitution	U	V	145	0	16	53	76	0
3	8	Substitution	L	LL	43	0	0	19	11	13
4	27	Substitution	V	U	42	42	0	0	0	0
5	19	Substitution	'	E	41	0	0	2	36	3
6	9	Insertion		'	35	0	16	0	19	0
7	25	Substitution	EE	E	34	0	16	1	0	17
8	2	Substitution	IE	Y	33	0	0	2	2	29
9	10	Insertion		E	26	1	2	8	5	10
10	26	Substitution	I	J	20	16	0	4	0	0
11	38	Substitution	S	SS	18	0	0	0	2	16
12	53	Substitution	I	E	16	0	0	1	15	0
13	42	Substitution	E	I	15	0	10	5	0	0
14	72	Substitution	Y	I	14	1	0	11	0	2
15	24	Substitution	TH	S	13	0	0	0	0	13
16	46	Substitution	LL	L	11	0	0	1	0	10
17	7	Deletion	-		11	0	1	10	0	0
18	67	Substitution	GG	G	10	0	0	4	6	0
19	57	Insertion			9	0	0	9	0	0

Fig. 3 Screenshot of DICER analysis on a manually standardized 5,000-word sample of Shakespeare's First Folio



Discovery and Investigation of Character Edit Rules

Shakespeare Sample

Rule #3: Delete E

Positions:

Position	Frequency
Start	0 (0.00%)
Second	0 (0.00%)
Middle	7 (1.81%)
Penultimate	53 (13.73%)
End	326 (84.46%)
Total	386

Character Groups:

Group	Frequency
Vowel Before	22 (5.70%)
Vowel(+Y) Before	26 (6.74%)
Vowel After	0 (0.00%)
Vowel(+Y) After	0 (0.00%)

Notes:

_ indicates a space character.

Previous Character:

Character	Total (%)+¹	Rule Position (%)			
		Second	Middle	Penultimate	End
K	17.36	0.00	0.00	19.40	76.12
N	16.32	0.00	0.00	9.52	88.89
R	14.77	0.00	0.00	8.77	87.72
L	12.18	0.00	0.00	6.38	93.62
S	7.25	0.00	0.00	0.00	100.00
F	5.96	0.00	0.00	0.00	100.00
O	5.44	0.00	0.00	4.76	95.24
D	5.18	0.00	0.00	10.00	90.00
M	3.89	0.00	0.00	26.67	73.33
P	3.11	0.00	0.00	8.33	91.67
G	2.33	0.00	0.00	44.44	55.56
T	2.07	0.00	0.00	12.50	87.50
W	2.07	0.00	0.00	100.00	0.00
Y	1.04	0.00	0.00	100.00	0.00
X	0.52	0.00	0.00	0.00	100.00
I	0.26	0.00	0.00	0.00	0.00

Next Character:

Character	Total (%)+¹	Rule Position (%)			
		Start	Second	Middle	Penultimate
S	13.73	0.00	0.00	0.00	100.00
-	0.78	0.00	0.00	100.00	0.00
L	0.52	0.00	0.00	100.00	0.00
R	0.26	0.00	0.00	100.00	0.00
T	0.26	0.00	0.00	100.00	0.00

Fig. 4 Screenshot of DICER showing the rule 'Delete E' in a manually standardized 5,000-word sample of Shakespeare's First Folio

By using DICER to analyze manually standardized samples of a corpus, a list of common character edit rules can be viewed. These character edit rules can then be added to VARD 2 and the tool will be better equipped to make judgments on the correct replacement for variants found whilst automatically standardizing the corpus.

In order to test VARD 2 and DICER's training ability a 5,000-word sample of Shakespeare's First Folio[6] was manually standardized in the interactive-mode of VARD 2 as training data, the entire corpus was then automatically standardized. Using this small amount of training data (6% of the entire corpus) increased the proportion of spelling variants replaced[7] from 70.33% to 73.75%.

The automatic standardization (after training) resulted in 10,601 unique variant replacements. 70.35% of these replacements could be achieved through VARD 2's original set of character edit rules alone. DICER analysis was then produced on the manually standardized Shakespeare sample; this is shown in Fig. 3. VARD 2's rule list was augmented with additional rules from the DICER analysis: any rule occurring 10 or more times was added, if not already present. Using this new rule list 77.66% of the 10,601 unique replacements could now be found, an increase of 7.31%.

The results are extremely promising, and increasing the size of the manually standardized sample should improve these figures even further. DICER can also be used to provide probabilities dictating how likely a character edit rule should be applied in a certain position with specified surrounding characters. Modifying VARD to use these probabilities could see even greater improvements in performance.

Calculating the precision of VARD 2's methods is difficult without a manually checked standardized corpus of decent size. We have recently acquired such a corpus and present the results of using this corpus to train and evaluate VARD 2's methods in Baron and Rayson (forthcoming).

Conclusion

This paper has described the problems that variant spellings cause for historical text mining, particularly for automated methods in historical corpus linguistics, such as part-of-speech tagging and key words analysis. In previous and forthcoming work, we have quantified the errors or differences that result from the application of untrained tools and techniques on historical data that has not been standardized. Our proposed solution is the VARD tool, which offers the potential to standardize spelling in historical texts automatically and with high accuracy. We have described recent improvements to VARD 2, such as the inclusion of a much larger modern dictionary that enables better detection of historical variants and matching with modern forms.

VARD 2 has been developed to deal with spelling variation in EModE texts; the tool can be used with its default settings to achieve partial standardization automatically. However, with some training, we have shown that VARD 2's performance is enhanced. Further training could allow the tool to be used with other varieties of non-standard English (e.g. SMS corpora and weblogs).

In the future, we will evaluate the extent to which variation that can only be detected contextually (e.g. 'then' for 'than' and 'bee' instead of 'be') contributes to the problem. Dealing with this problem requires more advanced techniques, e.g. POS tagging, to be used in the detection phase.

Notes

[1] Of course, spelling variants themselves are important linguistic features and thus worthy of study: as such, although our focus relates to how we might deal with spelling variation within historical data as a means of enabling the (more) effective use of automated analytical techniques, we advocate that any solution to this problem should always retain the original spelling. VARD 2 does so using an xml tag to note a replacement with the original spelling stored as an attribute of the tag.

[2] It should be noted that the accuracy of annotation is likely to be affected by additional factors, including differences in the grammar of the EmodE period when compared to present-day English (see Kytö and Voutilainen, 1995) and the possibility of a semantic shift in words from EModE to present-day English (see, for example, Knapp, 2000).

[3] The tool is available to download online, with a user guide also provided. The software is free to use for academic purposes from http://www.comp.lancs.ac.uk/~barona/vard2/

[4] Derived from the Spell Checking Oriented Word List (SCOWL). See http://wordlist.sourceforge.net/scowl-readme

[5] Available at http://juilland.comp.lancs.ac.uk/dicer/

[6] Available from the Oxford Text Archive: http://ota.ahds.ac.uk

[7] Variants here are words which VARD 2 deems to be variants, i.e. words which are not in its modern lexicon. It should be noted that words will be incorrectly marked as variants (particularly proper names) and some variants will be incorrectly marked as modern words (particularly *read-word errors*, such as 'bee' for 'be' and 'doe' for 'do').

References

Archer, D., McEnery, T., Rayson, P. and Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In Archer, D, Rayson, P., Wilson, A. and McEnery, T. (eds), *Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper number 16. UCREL, Lancaster University, pp. 22 - 31.

Baron, A. and Rayson, P. (forthcoming). Automatic standardization of texts with spelling variation, how much training data do you need? To appear in *Corpus Linguistics 2009*.

Baron, A., Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. In Ahrens, R. and Antor, H. (eds.) *Anglistik. International Journal of English Studies,* 20 (1), pp. 41-67.

Culpeper, J. and Archer, D. (forthcoming). The History of English Spelling. In Culpeper, J., Katamba, F., Kerswill, P., Wodak, R. and McEnery, T. (eds), *English Language and Linguistics*. Palgrave Macmillan, Basingstoke, UK.

Görlach, M. (1991). *Introduction to Early Modern English,* Cambridge University Press, Cambridge.

Knapp, P. A. (2000). *Time-Bound Words: Semantic and Social Economies from Chaucer's England to Shakespeare's*. Anthony Rowe Ltd, Chippenham, Wiltshire, UK.

Kytö, M. and Voutilainen, A. (1995). Applying the Constraint Grammar Parser of English to the Helsinki Corpus. *ICAME Journal* 19, pp. 23-48.

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In proceedings of *Corpus Linguistics 2007*, July 27-30, University of Birmingham, UK.

Rayson, P., Archer, D., Baron, A. and Smith, N. (2008). Travelling Through Time with Corpus Annotation Software. In Lewandowska-Tomaszczyk, B. (ed) *Corpus Linguistics, Computer Tools, and Applications – State of the Art*. *Palc 2007*. Peter Lang, Frankfurt am Main.

Vallins, G. H., and Scragg, D. G. (1965). *Spelling*. André Deutsch.

Generalizing the International Children's Digital Library

Benjamin B. Bederson
University of Maryland
www.cs.umd.edu/~bederson

Patrick Rutledge
University of Maryland

Alex Quinn
University of Maryland

Abstract

The International Children's Digital Library is a website of exemplary free children's books from around the world (www.childrenslibrary.org). Since we launched the website in 2002, we have recently embarked on an effort to generalize what we have learned, creating a version (to be made open source when it is finalized) that is suitable for more general content and a more traditional adult audience. We also have been exploring mobile deployment – providing access to the ICDL's picture books with an interface that displays readable text in context on Apple's iPhone. Together, these efforts demonstrate how we can learn from one successful project to expand into other areas of content and platform, and in this paper we attempt to summarize the core lessons learned that can be applied elsewhere.

Introduction

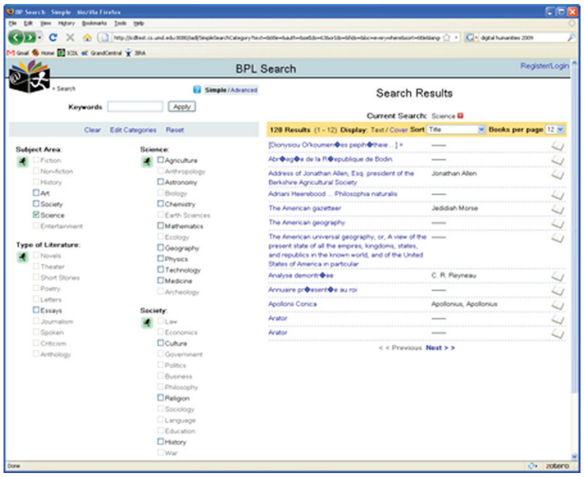
The International Children's Digital Library (ICDL) is an established electronic archive, providing children and their adults easy access to thousands of children's books in almost 50 languages which can be read online for free. Over the six years since we deployed the ICDL, we have learned a lot about what it takes to make a usable digital library (Bederson 2008). Motivated by a need of the Boston Public Library and the Knowledge Commons (formerly the Open Content Alliance) to offer a highly usable range of interfaces for the millions of books they are scanning, and by the recognition that people are accessing more and more books on mobile devices, we decided to start branching out, and applying those lessons to other content domains, audiences, and platforms.

Content

The books in the ICDL are all digitized versions of traditionally printed paper books which are selected for children ages 3-13. Almost half are picture books. Due to the narrow range of content, we have been able to manually catalog the books in the library with a fixed ontology of just under 300 categories in a hierarchical structure (i.e., Appearance->Format->Picture Books). We use a combination of traditional categories (such as "Short Story") and more customized categories that we developed in the course of working with children (such as "Book cover color" and whether the books are "Happy" or "Sad").

In order to design the ICDL interface to support a broader set of content, we had to rethink how we categorized the books and presented those categories. A few hundred categories do not offer a fine enough distinction for millions of books (Baker 1996). In fact, the standard MARC records of the Library of Congress use many thousands of categories. The traditional solution is remove the focus on categories, and to make textual search the primary mechanism for searching for books in large collections. However, this gives up the power of categorization, especially when the person has some particular attributes they want to specify, or if, for example, a teacher wants to specify a set of categories that might be useful for their students to use for more focused searching. Another scaling problem with our original design is that we used manually created icons for each category. This was a lot of work for a few hundred categories, but impossible when there are many more.

We came up with a new solution (Figure 1) that lets the user choose which categories are displayed in the primary search window and integrates keyword search, the search results, and a preview of the book. A tightly coupled advanced search (not shown) adds several features, including a more hierarchical display of categories, counts of how many books are in each category, explicit search fields for title & author, and more.



[Dionysiou Oi'koumen◆es pepih◆theie ...] »	——
Abr◆eg◆e de la R◆epublique de Bodin	——
Address of Jonathan Allen, Esq. president of the Berkshire Agricultural Society	Jonathan Allen
Adriani Heerebood ... Philosophia naturalis	——
The American gazetteer	Jedidiah Morse
The American geography	——
The American universal geography, or, A view of the present state of all the empires, kingdoms, states, and republics in the known world, and of the United States of America in particular	——
Analyse demontr◆ee	C. R. Reyneau
Annuaire pr◆esent◆e au roi	——
Apolloni Conica	Apollonius, Apollonius
Arator	——
Arator	——

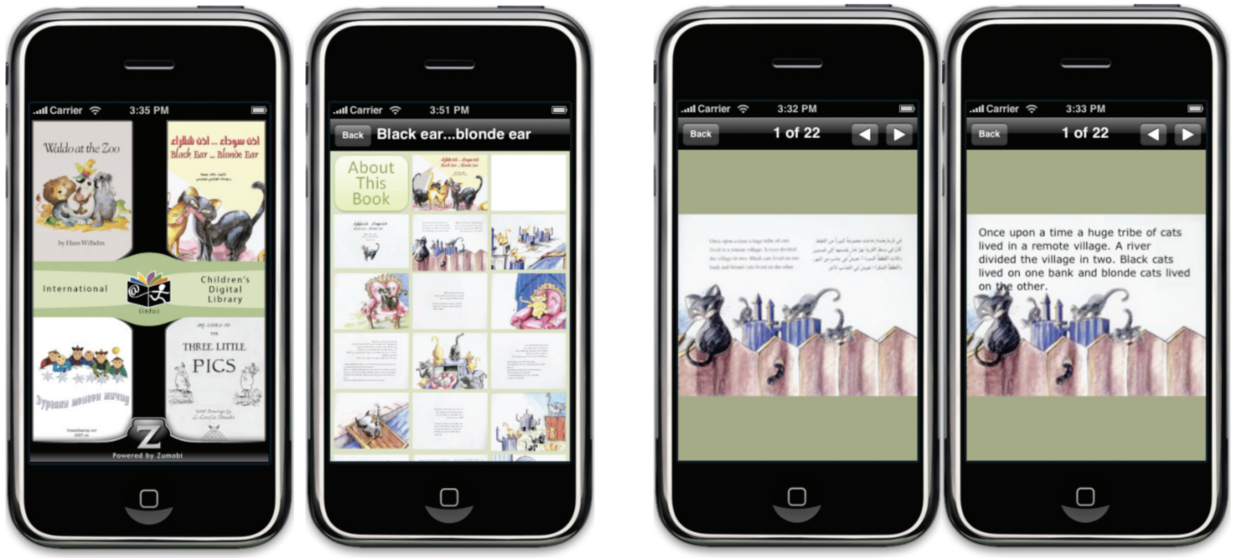
Figure 1: Current prototype for generalized version of ICDL For broader content and older users.

Audience

Our aim of supporting adult users and not just children also gave us some flexibility in designing this prototype. Our research with children has led us to understand the importance of using large objects to click on the screen, fewer abstractions in the interface and a generally simpler visual display (Hutchinson 2003). The interface we created thus correspondingly uses more text, smaller hit targets, and a visually denser display.

Platforms

In the same spirit of trying to make books available to more people, we designed a version of ICDL for the Apple iPhone platform. This platform is unusual because while small, it has a relatively high resolution display and powerful graphics that enable smooth animated transitions. We built an application that gives offline access to four picture books. It uses the "ClearText" technology that ICDL developed for making text legible on small displays (Quinn et al. 2008) and smooth animated transitions that make it easier to understand where you are in the interface and book as the reader navigates a complex information space on a very small display.



Once upon a time a huge tribe of cats lived in a remote village. A river divided the village in two. Black cats lived on one bank and blonde cats lived on the other.

Figure 2. (Left to right): a) ICDL for iPhone home screen shows the entry point to 4 books for iPhone and iPod Touch platforms; b) the result of zooming into the top-right book (Blond ear … black ear); c) the result of zooming into a page of that book, and the result of tapping on a textbox to display it in a font that is large enough to read.

Concluding Thoughts

As the trends of increasing technological capabilities and broad increases in access continue to occur, we must refine and innovate the way we deliver access to books. It is crucial that people have an unfettered ability to find what they are looking for, and to deeply engage with the content of books once they have found it. The efforts described here show two approaches to achieving this goal.

The lessons from the web interface include the importance of building an all-in-one interface with varying levels of search complexity along with search results and book previews all integrated into a single screen. A related lesson is that you can't ignore issues of scale in your design, and that one important approach is to couple a very good and simple initial experience with the possibility of significant end-user customization.

The lessons from the iPhone interface are that you must re-think what you are offering for this tiny and mobile package. You cannot and should not do everything that you offer on your full website. It was a tough call for us, but we completely eliminated search! The other issue is that while difficult, you must balance engagement and usability. We tried to use interesting animations to keep the experience playful while not getting in the way of reading – and we also spent a lot of time creating an interface that supports the core mission of the application – which is clear and legible text for easy reading. As the application is just being released now, time will tell whether we were successful with these goals.

References

Baker, S. (1996). "A Decade's Worth of Research on Browsing Fiction Collections."

In Kenneth Shearer (Ed), *Guiding the reader to the next* *book*. New York, NY: Neal-Schuman Publishers, Inc. 45-72.

Bederson, B.B. (2008) “Experience the International Children’s Digital Library”, *Interactions Magazine*, ACM Press, 50-54.

Hutchinson, Hillary Brown (2003). Children’s Interface Design for Hierarchical Search and Browse. In proceedings of *ACM SIGCAPH Computers and the Physically Handicapped*. 75: 11-12

Quinn, A., Hu, C., Arisaka, T., & Bederson, B.B. (2008) Readability of Scanned Books in Digital Libraries, in proceedings of *ACM CHI (CHI 2008)*, ACM Press, 705-714.

Snake’s Nest: Untangling the Relationships between Classic Maya States

Alex Bennett
University of Portsmouth
Alex.Bennett@port.ac.uk

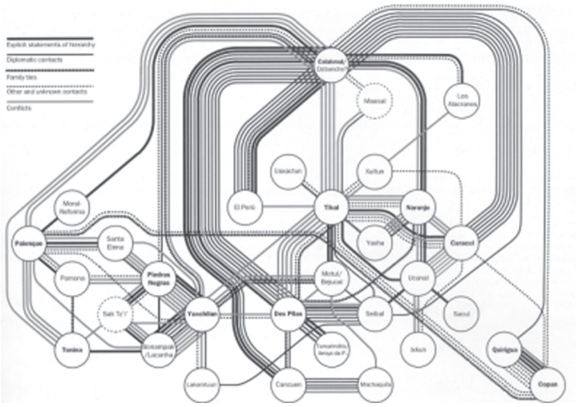
The political organisation of the Classic Maya world has generated almost as much conflict and contention among scholars as it did amongst the ancient Maya themselves. As Prudence Rice [2004] has amply demonstrated, prevailing theories have swung from a homogeneous empire to a fractious patchwork of independent statelets existing in precarious balance and most conceivable organisational models between. This article discusses the the problems in trying to untangle the complex inter-state relations of the Classic period and the development of a software tool to support further research in this area.

The ancient Maya have a complex historiographical relationship with the modern world. Without wheeled transport or metal tools they developed an intricate and sophisticated culture that after its disintegration around 900AD vanished almost completely back into the forests: until the middle of the nineteenth century hardly anyone in the western world knew much more than the wild tales of shining cities buried in the jungle and the search for El Dorado. The ancient Maya came to popular attention at the height of the age of reason, as archaeology was evolving from an aristocratic pastime into a science. In an age that equated civilisation with literacy, the Classic Maya were all but mute. Early explorers were sure their monuments recorded written information but the script remained largely undeciphered. An understanding of the calendrical system allowed early historians to develop a broad chronology but in the absence of any meaningful testimony from the Maya themselves, imagination was employed to fill the blanks.

Many of these gaps were filled by forcing the Maya to conform to the social and political templates provided by better understood ancient cultures, with limited success. The emphasis on dates in the epigraphic record fuelled an optimistically utopian vision of a pacifist people living in splendid isolation under the rule of benevolent astronomer priests obsessed with the cycles of time. Its most famous advocate Eric Thompson [Drew 1999] acknowledged evidence of conflict but dismissed it as little more than minor border disputes or contested inheritances: to follow the argument to its logical conclusion, these events might have been recorded for rarity value.

Breakthroughs in decipherment made the utopian ideal increasingly untenable. The emerging epigraphic record presented a tangled, often bloody history of warrior rulers, holy lords recording their achievements in birth, alliance, conquest and death in a medium designed to awe their subjects and cow their enemies. Recent research [Martin & Grube 2000] has identified glyphs for hierarchical relationships amongst rulers suggesting a web of coalition and patronage, covering vast distances, that may represent anything from an ephemeral expression of dominance to a long-term relationship such as that attested to between Tikal and Palenque.

The richest and most widely reproduced diagrammatic representation of this political structure comes from Simon Martin [Martin & Grube 2000, p21]. It demonstrates that, for much of the Classic period, the rivalry between the polities of Calakmul and Tikal formed the nexus around which the wider political system revolved.



Known in Classical times as Kaan – the Snake Kingdom – Calakmul was a giant by the standards of the ancient world, boasting an urban population estimated by population density analysis to be around 90,000 with a suburban rural population of over 2.5 million under its immediate authority [Braswell *et al* in Demarest, Rice & Rice 2004]. It is the most frequently referenced site on monuments throughout the region and has been labelled a 'super state' or, rather more emotively, a superpower [Coe 1967; Martin & Grube 2000 *inter alia*]. Unlike other powerful states such as Tikal and Palenque, the quality of local stone at Calakmul has meant that many inscriptions are now fragmentary or illegible: the majority of what is known about Calakmul's political machinations comes from other sites. The prevailing impression, coloured perhaps by Western conceptions of the snake's character, is of an acquisitive and aggressive evil empire, overcome by the dogged resistance of Tikal and her allies.

There is no doubting this interpretation could be accurate: epigraphic evidence and the Martin diagram clearly show Calakmul was politically and militarily active throughout the Maya world. The Martin diagram may, however, be contributing to an impression of continuous interference over the entire span of the Classic period that cannot be supported by the fluid nature of political relations in the Maya world demonstrated elsewhere. In a social system where power and prestige are directly influenced by the strength and even charisma of the ruling lord, many inter-state relationships would come and go in a handful of years.

The strength of the Martin diagram is its ability to summarise a wealth of detail about the relationships between the largest Maya polities. In many ways, it resembles the network map of the London Underground, with the same advantages and attendant weaknesses. While it distinguishes types of association, for example conflict from diplomacy, and gives some indication of the frequency of those links, it is unable to demonstrate their duration or the evidence. A conflict link, for example, may represent a single act of aggression like a religiously sanctioned Star War or sporadic low-intensity warfare over an extended period. In cases such as the relationship between Calakmul and Yaxchilan which shows both conflict and marriage alliance, it is impossible to distinguish which came first or indeed whether they were related at all: the marriage attested to between Bird Jaguar of Yaxchilan and Lady Evening Star of Calakmul [Drew 1999] could have been arranged to cement peace or the two events could be generations apart.

The very short period between 680 and 685 AD provides a good example of the complexity the Martin diagram is poorly equipped to show. Those five years saw a dramatic shift in the fortunes of many of the major powers across the Maya world: Calakmul, Palenque and Dos Pilas all lost rulers who had been in power for over forty years, notably long reigns even by modern standards; in contrast Tikal and Naranjo were both emerging from periods of weakness and internal division. Piecing together the direct effects of these changes on individual polities has proved difficult enough but recognising the effects on the wider political landscape is almost impossible with a diagrammatic model that cannot distinguish change over time.

The aim of this research is to produce a new interaction mapping tool capable of overcoming some of these weaknesses in the Martin diagram. The Dynamite (*Dynamic* *M*aya *I*nformation *T*ool) software is designed to produce a rich, updateable model of political interactions that can be extended and customised to support the needs of individual researchers. In its standard format, it provides an enhanced version of the Martin diagram, allowing users to filter the data on a range of different criteria, including regionally and temporally. It can also support more specialised filtering to show interactions and source evidence by type: employing the full range of filters it would be possible to examine nothing but Calakmul's military conquests for the period 580 to 620 AD, derived solely from epigraphic sources.

Dynamite uses a model of loose data connectivity based on XML structured data to provide maximum flexibility and extensibility. Beyond compatible data typing, it makes no assumptions about the relationship between objects: all existing filters are provided through stored queries. This allows researchers to be flexible and innovative in how they use and expand the source data while the XML format supports the easy sharing of results and helps Dynamite move towards a structure of plug-and-play data.

The suite of Dynamite support tools allows researchers with even minimal computing experience to examine, update and extend the stored data as well as create new queries to run against that data. There are also shortcuts for experienced developers providing more direct access.

It is hoped that the Dynamite software will contribute positively to the next generation of research tools for exploring the growing body of archaeological and epigraphic data on the relationships between Maya polities of the Classic period.

Bibliography

Aoyama, K [2005] *Classic Maya Warfare and Weapons: Spear, Dart and Arrow Points of Aguateca and Copan*; Ancient Mesoamerica 16

Boot, E [2002] *The Life and Times of Balah Chan Kawil of Mutal According to Dos Pilas Heiroglyphic Stairway 2*; MesoWeb

Coe, M [1967] *The Maya 7th Edition 2005*; Thames and Hudson

Coe, M [1992] *Breaking the Maya Code 2nd Edition 1999*; Thames and Hudson

Demarest, A, Rice & Rice (eds.) [2004] *The Terminal Classic in the Maya Lowlands: Collapse, Transition and Transformation*; University Press of Colorado

Drew, D [1999] *The Lost Chronicles of the Maya Kings*; Orion Books

Johnston, K [2001] *Broken Fingers: Classic Maya scribe capture and polity Consolidation*; Antiquity Volume 75

Kistler, A [2004] *The Search for Five-Flower Mountain: Re-evaluating the Cancuen Panel*; MesoWeb Online Articles

Martin, S [2005] *Of Snakes and Bats: Shifting Identities at Calakmul*; PARI Online Publications

Martin, S and Grube, N [2000] *Chronicle of the Maya Kings and Queens: Deciphering the Dynasties of the Ancient Maya* Revised Edition 2008; Thames and Hudson

Rice, P [2004] *Maya Political Science: Time, Astronomy and Cosmos* University of Texas Press

Stuart, D [2004] *The Paw Stone: The Place Name of Piedras Negras, Guatemala* PARI Online Publications

Clustering the Short Stories of Edgar Allan Poe Using Word Groups and Formal Concept Analysis

Roger Bilisoly
Central Connecticut State University
bilisolyr@ccsu.edu

The proposed poster will summarize recent work on finding clusters of Edgar Allan Poe's short stories. These are defined by high rates of word usage for five word collections, each of which is defined by a similar meaning. For example, one of them focuses on *death* and has terms such as *corpse*, *dead* and *die*. By choosing word groups pertinent to the types of stories that Poe wrote, the resulting clusters make more intuitive sense than other clustering algorithms that are common in information retrieval.

Humans enjoy dividing texts into similar clusters. For example, the literary critic Daniel Hoffman discusses Poe's stories in thematic groups (Hoffman, 1990). For instance, vendors such as Amazon.com create clusters that consist of recommendations made to users, which are based on past purchases. Consequently, there is value in creating a clustering algorithm that can explain how the grouped texts are similar in terms that a human can appreciate.

Unfortunately, the most popular techniques in information retrieval for computing text similarity produce results that lack appeal to human sensibilities. For example, the vector space model (Section 2.1 of Grossman and Frieder, 2004) reduces a text to a table of (usually weighted) word counts where each column represents a text and each row represents a word. Each column can be thought of as a vector, so the geometric idea of angles between texts can be introduced, where smaller angles correspond to higher similarity. This geometric point of view is powerful because it is well understood by mathematicians, and it already has been applied to studying information. For example, this is the focus of the book *Geometry and Meaning* (Widdows, 2002). However, the dimensions of these vectors can be quite high (in the tens of thousands), which makes the results hard to reconcile with human intuition.

The technique used in this poster builds on the vector space model as follows. Instead of including all the words that Edgar Allan Poe uses in his short stories (about 20,000 total), these are analyzed to find a few clusters, each defined by a shared meaning, which are built by using a thesaurus. Note that this removes function words such as *the*, *of*, *and*, *a* and *to*, which primarily serve grammatical roles and are less interesting to a human reader. Five groups are used in this poster, which are based on the following themes: death, body, spiritual, horror, and family. For example, the top ten most frequent death words in Poe are *death*, *corpse*, *dead*, *murder*, *died*, *die*, *deceased*, *dying*, *fatal*, and *deadly*. Anyone familiar with his short stories would not be surprised by these word groups. For example, many of his stories are about people who die (for instance, "Morella," "Eleonora," and "The Oval Portrait") or are killed (for instance, "The Black Cat," "The Murders in the Rue Morgue," and "The Tell-Tale Heart").

Since word frequencies are a function of text length, Poe's story lengths must be taken into account. This can be done with word rates, that is, the proportion of a word in a text. Unfortunately, these can also depend on text length (see discussion in Chapter 1 of Baayen, 2002), so stories are grouped by length, which are then analyzed separately. Three groups are used in this poster: 2000 to 3000, 3001 to 4200, and 4201 to 6000.

For each of these three ranges, matrices are constructed where each row represents a word group, each column represents a Poe short story, and the entries are word group rates. Although the vector space model could be used in this situation, the alternative method of formal concepts is used because it is more interpretable. Formal concept analysis (FCA) is a technique to create a double lattice of concepts given a list of objects and another one of attributes. It is a central tool in concept data analysis (Carpineto and Romano, 2004) and is used to organize information in a way more similar to humans. For this application, the objects are Poe's short stories, and the attributes are a story having a word group rate above the median.

A formal concept consists of a set of objects and a set of attributes, where each object shares all the attributes, and each attribute is shared by all the objects. This set of objects is a cluster, which is describable by its shared attributes. Applied to Poe's short stories, each cluster is determined by its use of the five word groups defined above. Since a list of *death* words is both straightforward to compute and is evocative for a human reader, such story clusters have intuitive appeal. Moreover, there are several algorithms published that efficiently find all the formal concepts. For this project Ganter's algorithm is used (Ganter, 1984).

Here is a specific example. There are thirteen Poe stories that are between 2000 and 3000 words long. We use the five attributes defined above, one for each word group. A story has one of these if its word group rate is above the median rate for all thirteen stories. For example, the median death word rate is 1.22 words per thousand (wpt). The six stories with higher rates have the attribute *death*, the other seven do not. For body words, the median is 5.44 wpt, so the six stories above this threshold have the *body* attribute. The same computation is done for the remaining three word groups.

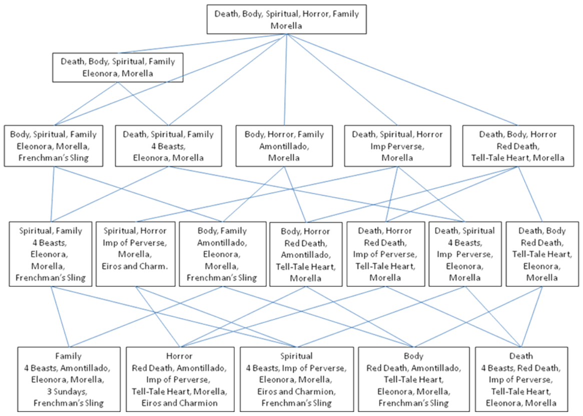


Figure 1. Formal concept lattice for Edgar Allan Poe stories with lengths between 2000 and 3000 words.

Formal concepts always form two lattices, called a Galois lattice. One focuses on the stories, and the other focuses on attributes. Both of these are ordered by subsets, and these are related: as the number of attributes increases, the number of objects decreases, and *vice versa*. In this application, nineteen formal concepts are found, which are given in Fig. 1. We consider one here, which consists of the Poe stories "The Tell-Tale Heart," "The Masque of the Red Death," and "Morella," and the attributes *death*, *body*, and *horror*. Although the plots of these stories differ, there are similarities apparent to a human. First, all three stories are about death. In "The Tell-Tale Heart" the narrator tells how he kills his roommate. The red death is a pestilence, and in "Morella" the narrator tells of his wife's obsession with death and rebirth. Second, all three stories discuss the body. The first one features an evil eye and a beating heart, while the *red* of the second one refers to blood. Morella dies giving birth, and her daughter grows to be just like her mother physically, which is described in the story. Finally, it is no surprise that stories about death that include bodily descriptions would have many words related to *horror*.

The tools of FCA and Galois lattices used in this poster are flexible. It is this author's belief that additional applications will be found, both for other literature and for other types of texts.

References

Baayen, R. H. (2002). *Word Frequency Distributions*. New York: Springer.

Carpineto, C. and Romano, G. (2004). *Concept Data Analysis: Theory and Applications*. Chichester: Wiley.

Ganter, B. (1984). Two basic algorithms in concept analysis. Technical Report FB4 – Preprint No. 831, TU Darmstadt, Germany.

Grossman, D. A. and Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics*, 2nd Edition. Dordrecht: Springer.

Hoffman, D. (1990). *Poe Poe Poe Poe Poe Poe Poe*. New York: Marlowe & Co.

Widdows, D. (2004). *Geometry and Meaning*. Palo Alto: Center for the Study of Language and Information.

"Song(s) of Myself": Flexing *Leaves of Grass*

Olin Bjork
University of Texas, Austin
olin.bjork@gmail.com

Scott Herrick

In his monograph *Radiant Textuality*, Jerome McGann reflects on the first decade of work on the *Rossetti Archive* at the Institute for Advanced Technology in the Humanities and regrets their "failure to consider interface in a serious way….when we worked out the archive's original design, we deliberately chose to focus on the logical structure and to set aside any thought about the Interface for delivering the archive to its users. We made this decision in order to avoid committing ourselves to a delivery mechanism."[1] Accordingly, the site's design editor, Bethany Nowviskie, predicts that the interface will "always have something of a tacked-on quality."[2] Such an outcome, Matthew Kirschenbaum argues, is consistent with the standard workflow of Digital Humanities projects: "Too often put together as the final phase of a project under a tight deadline and an even tighter budget, the interface becomes the first and in most respects the exclusive experience of the project for its end users."[3] This assessment suggests that future editors and directors should consider making interface design a preliminary stage in the process of constructing electronic editions and archives.

In 2004, Professor John Rumrich and myself, then a graduate student at the University of Texas at Austin (UT-Austin), decided to reshuffle the priorities of the Digital Humanities in two ways: by considering interface design before text encoding and by privileging pedagogical applications over scholarly ones. Noting that most electronic editions and archives have either deliberately or unreflectively adopted the standard windows-based interface design of non-literary Web sites, we resolved to make our digital classroom edition of John Milton's *Paradise Lost* (hereafter, *PL*) resemble a book lying open on a table. This approach, we hoped, would increase readability for students and appeal to bibliophile instructors. But we refused to rely solely on visual usability—we wanted to integrate an audio track of the poem with our electronic text. To this end, we recruited several colleagues at UT-Austin to make a recording of one of the epic's twelve books. From 2005 to 2007, with the support of a UT-Austin Liberal Arts Instructional Technology Services (LAITS) grant, we developed a prototype "audiotext," which we demoed at the 2007 Digital Humanities conference (DH07). In our abstract, we offered the following rationale for the edition's synthesis of audio and text:

> Whether students run through excerpts from *PL* in a sophomore survey or pore over the entire epic in an upper-division course, they famously find Milton's poetry difficult to follow. Instructors usually assume that this difficulty owes to its unfamiliar ideas and Milton's intimidating erudition—and surely these are part of the problem. But we have found that when students hear an instructor declaim passages from *PL* as they follow along in their textbooks, the thrust of the lines suddenly becomes plainer. Recent research on multimedia learning indicates that distinct, additive cognitive pathways mediate the aural and visual reception of language (Mayer 2001). Reading and listening to the same text demonstrably improves understanding and recall.[4]

The prototype uses Adobe Flash technology to synchronize the audio playback with a karaoke-style highlight that moves over each line as it is spoken. Since the conference, two more books have been added to the project Web site at http://www.laits.utextas.edu/miltonpl. The *PL* audiotexts are used by several instructors each semester in a range of British literature courses at UT-Austin. Student survey data and instructor comments have been overwhelmingly positive.

In 2007, Professors Coleman Hutchison and Michael Winship, Americanists intrigued by the interface concept, applied for and received an LAITS grant to create an audiotext of Walt Whitman's "Song of Myself," arguably the best-known long-form poem in American literature. The rationale they stated in their grant proposal was not overcoming a syntactical barrier to comprehension so much as countering the problem that Whitman's long, idiosyncratic lines and extensive use of poetic catalogs can *look* like prose. Hearing as well as seeing the text, Hutchison and Winship argued, should help students gain an understanding of the aural register of the poem and thereby become receptive to the concatenation of Whitman's aural and visual effects. The editors also proposed that six different readers would each voice the entire poem, whereas in the PL audiotexts each character in the poem is voiced by a different reader. The project would thus highlight an understudied aspect of the poem: despite its title, "Song of Myself" is a polyvocal poem, one in which distinct personae and voices compete. Furthermore, by offering a menu of six audio tracks from which a user can select at any time, changing tracks without interruption, the new interface would call attention to the subjective nature of such vocalization factors as accent, pronunciation, and emphasis. This “remixability” addresses an issue raised by otherwise enthusiastic listeners/observers at DH07 and elsewhere concerning the *PL* prototype. Although Professor Rumrich and I considered our audio track to be no more “accurate” a rendering of Milton’s lines than the visual representations we included in the illustration section, we overlooked the possibility that a single audio recording would give the impression that we were claiming to have realized Milton’s auditory intentions, just as a single critically edited text strikes some textual critics as more autocratic than an archive of versions.

Hutchison and Winship hired me, now a postdoctoral fellow at Georgia Tech, as the project’s technical editor. The project manager, Emily Cicchini, hired Scott Herrick, a systems analyst for the Division of Instructional Innovation and Assessment at UT-Austin, as lead developer. Herrick had experience constructing a streaming flash architecture for the *Aswaat Arabiyya* project (http://www.laits.utexas.edu/aswaat/). In order to minimize the time necessary to load six audio tracks, we decided to stream the audio tracks from a Flash Server instead of embedding the audio in the Flash movie, as had been done in the PL prototype. Herrick and I were aware that the choice of Flash as the core technology might prove a liability for the project’s long term sustainability as well as its adaptability to other contexts and purposes, but we had developed a familiarity with the platform and were not confident that we could achieve the desired animation and synchronization effects using open-source alternatives. Although the Flash Player is a free download, it is proprietary to Adobe, whose standard Integrated Development Environment (IDE) for Flash is a closed, commercial product. Adobe released the specifications for its playable Flash formats (SWF and FLV) in 2006, a move that some believe was intended to discourage the development of a free and open-source alternative while increasing the number of third party applications capable of generating SWF files.

Lately, Adobe has been promoting Flex Builder, an alternative Flash IDE built on the open source Eclipse platform yet itself proprietary. The new IDE resembles a Java development suite and combines pre-built application components with MXML, an XML-based markup language, and ActionScript 3.0, an object-oriented scripting language. Flex Builder is designed to appeal to programmers and designers of Rich Internet Applications who were never comfortable with the animation and movie metaphors of the timeline-based standard Flash IDE. Noting that Adobe is currently offering Flex Builder for free to educators, Herrick suggested that we build the “Song of Myself” audiotext in that environment. To minimize the constraints of Flash, we separated all of our content from the interface. The audio tracks stream from the server as needed, while the text is imported into the interface from a TEI (Text Encoding Initiative) XML file at runtime. Suloni Robertson, an LAITS graphic designer and artist, executed the background image, which resembles an opened copy of the original edition of *Leaves of Grass*, published in 1855.

The poem that would later be titled “Song of Myself” opens this volume, and the beta of our audiotext version is now available at http://www.laits.utexas.edu/leavesofgrass. In addition to the audio controls, the application includes a zoom function and an options panel with a search engine and selection tools. In the next phase of the project, we will add explanatory notes and an animated page-turn effect. Ultimately, we plan to include a full transcript, critical apparatus, and facsimile images of the entire book. At that point, the audiotext will become a full-fledged scholarly edition without compromising its pedagogical utility or its innovative interface design. The editors of the *Walt Whitman Archive* have generously allowed us to use their transcript from a copy of the 1855 edition (http://www.whitmanarchive.org/published/LG/1855/whole.html)_as the basis for our electronic text. Professor Hutchison has had conversations with these editors about adapting the audiotext concept to create a more robust front-end option for the archive’s resources. To this end, we are exploring the possibility of using an open-source alternative to the Flex IDE to create a set of audiotext templates that could be used by other projects.

Notes

[1] Jerome McGann, *Radiant Textuality: Literature after the World Wide Web* (New York: Palgrave, 2001), 141.

[2] Bethany Nowviskie, “Interfacing the Rossetti Hypermedia Archive,” (paper, Humanities and Technology Association Conference, Charlottesville, VA, October 2000), http://www.iath.virginia.edu/~bpn2f/1866/dgrinterface.html.

[3] Matthew G. Kirschenbaum, “‘So the Colors Cover the Wires’: Interface, Aesthetics, and Usability,” *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth (Oxford: Blackwell, 2004), 525.

[4] Olin R. Bjork and John P. Rumrich, “The *Paradise Lost* Flash Audiotext” (poster, ACH/LCC Digital Humanities conference, Champaign-Urbana, IL, June 2-8, 2007), http://www.digitalhumanities.org/dh2007abstracts/xhtml.xq?id=216.

Can Pliny be one of the muses? How Pliny could support scholarly writing

John Bradley
King's College London
john.bradley@kcl.ac.uk

In Bradley 2008a the software *Pliny* (Pliny 2007) is described as a tool to support traditional scholarship, which, in turn, is assumed to be based on the reading of primary and secondary texts and the eventual writing of new secondary texts that describe an interpretation that has emerged in the scholar's mind as s/he worked with his/her materials (p. 266). In that paper I described mechanisms – there called *affordances* – that Pliny provides to support personal annotation and the use of personal annotation to support the development of an interpretation.

Fig. 1 (taken from my presentation at the DH2008 conference: Bradley 2008b) shows, in a way somewhat different from that shown in Bradley 2008a, where Pliny is meant to fit in the three-stage processed described in Bradley 2008a.

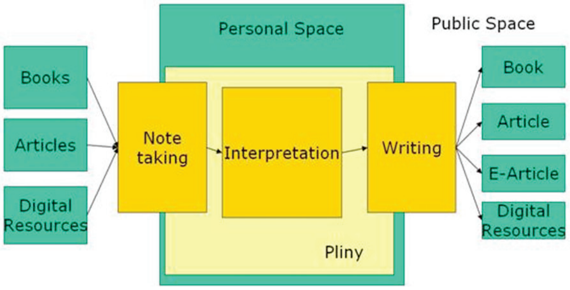


Fig. 1: Pliny and a scholar's personal space

Perhaps this figure makes it more evident that Pliny fits between two borders that separate public and personal space: one that takes non-personal materials as inspiration in (on the left) and a second one (on the right) that puts personal materials out into the public domain again in the form of a book or article. In past work on Pliny I have focused primarily on the boundary on the left in Fig. 1 and on how the computer might support the development of a personal interpretation (in the middle). For this poster I'd like to extend some thinking, and promote some thinking by others, about the boundary between the public and personal space that is shown on the right – output *from* Pliny rather than input *into* it.

Section 6 of Bradley 2008a described the beginnings of a kind of strategy for how to use Pliny's interpretation development affordances to support the writing of a traditional presentation. It also touches on the way that Pliny can export its formal model of the materials it holds about an interpretation into a computer-ontological-kind of format – currently Topic Maps (Biezunski et al., 2002). There is certainly interesting further work to do to explore the potential of the formal output of Pliny materials to represent scholarly thinking in, say, XML, but this is not the focus of this poster. Instead, I would like to focus on strategies that could be built into Pliny to assist its user in the transformation of materials that s/he has put into Pliny into a *traditional* prose text.

The task of putting ones thoughts into the linear order of a text that is understandable to others is quite evidently a difficult one, and although a 2D Pliny-like representation of the things one wishes to discuss is helpful, it seems, by providing a holistic personal overview of your materials that could be used to guide writing, the task of taking this overview and transforming it into prose is still difficult. There seem to be two issues: one is captured by the seeming difficulty of fully representing a 2D hierarchical Pliny reference space as a 1D (temporal) hierarchical object that is, superficially at least, the basis of scholarly writing. The second relates to what happens to personal materials such as the kind of notes one makes in Pliny – insights perhaps – when they become public objects. Catherine Marshall (Marshall 1998, pp. 40-42, and again in Marshall and Brush 2004) categories personal annotation, for example, as different from public annotation – more informal, more terse, likely often even enigmatic to the outside reader – and it is in this difference that perhaps the challenge of transforming a collection of these things into prose meant to be read by others also resides.

There is, of course, a substantial amount of work that theorises about the act of writing and reading. I confess that I am not in a position to fully take all of this in, although among my readings in this area I have found the work of George Landow (see, for example, Landow 1997), to be as close as anything else I have read that tries to blend theoretical and practical in ways that provide useful pragmatic clues about the process. My aim with Pliny is primarily practical rather than theoretical, although I would wish, where possible, to do work here that would be informed by whatever practical insights can be drawn from the theorising that has been done about scholarly writing. An alternate, and definitely more practical stream of thinking comes from computer science in the work of Marshall, Frank Shipman and others on software like VIKI and VITE and with social sciences theorists in their thinking about tools to support qualitative analysis (see Shipman *et al* 1995, and Hsieh and Shipman 2000 for a computer science perspective and Pandit 1996 for an overview of that from the social sciences). In between, perhaps, is the work of Linda Flower, who in Flower 1988 recognises the rhetorical element in the act of transforming ideas into prose. Much of Flower's article is based around the transformation of what she calls the "writer's web of purpose" (pg 532-534): a verbal image that seems compatible to the graph-oriented model for ideas that Pliny supports.

Work in Pliny so far in this area has been influenced mostly by some of Marshall and Shipman's work and has centered on recognising the challenges inherent in taking a 2-D Pliny representation of an interpretation and expressing it satisfactorily in the seeming essentially 1-D temporal context of scholarly writing. The Pliny user can, of course, create an object specifically to represent a piece s/he is writing, assemble issues of interest and play around with them to find a set of relationships between them that work best, and then use the resulting 2-D space to think about a 1-D ordering that will guide her/his writing. Pliny can currently look for and exploit in a rudimentary way such visual structures that the user has created to guess how the space might be best mapped into one dimension, and work is now underway to enhance Pliny's ability in this area through applying some of the facilities described in writings about VIKI. Some work has been done as well to assist the user in managing the mix of public and private objects in his/her Pliny connection so that ones aimed at the public are exported to be taken up in the writing of an article.

The questions to be asked, then, are (a) what are the intellectual challenges to taking a holistic 2D model of an interpretation of the kind one can build in Pliny and turning it into a text, and can Pliny assist this in ways better than it does now, and (b) what issues arise in the transformation of private materials such as one accumulated in Pliny into a public text, and can Pliny help there more than it does now. I'd be delighted to hear ideas from conference attendees.

References

Biezunski, M., Newcomb, S., and Pepper, S. (ed.) (2002). ISO/IEC 13250 Topic Maps (2nd edn). International Organization for Standardization. Online at http://www.1.y12.doe.gov/capabilities/sgml/sc34/document/0322_files/iso13250-2nd-ed-v2.pdf (accessed 7 Novemberr 2008).

Bradley, John (2008a). Thinking about Interpretation: *Pliny* and Scholarship in the Humanities. In *Literary and Linguistic Computing* Vol. 23 No, 3, 2008, pp. 263-79. doi: 10.1093/llc/fqn021. Online at http://llc.oxfordjournals.org/cgi/reprint/fqn021?ijkey=3UzJDubDB0FRQcR&keytype=ref . (accessed 7 November 2008).

Bradley, John (2008b), "Playing together: modular tools and Pliny". Peer reviewed paper given in the Digital Humanities 2008 conference, University of Oulu, Oulu, Finland, 28 June 2008. A draft version of this is online at http://pliny.cch.kcl.ac.uk/docs/oulu-paper.html. Accessed 7 November 2008.

Flower, Linda (1988). "The Construction of Purpose in Writing and Reading". In *College English*, Vol. 50. No. 5 9Sept. 1988). pp. 528-550.

Hsieh, Hao-wei and Frank M. Shipman III (2000). "Vite: A Visual Interface Supporting the Direct Manipulation of Structured Data Using Two-Way Mappings". In Procedings of the IUI Conference, New Orleans. pp. 141-48.

Landow, George P. (1997). *Hypertext 2.0: Being a Revised, Expanded Edition of Hypertext: the Convergence of Contemporary Critical Theory and Technology*. Baltimore, MD: John Hopkins University Press.

Marshall, Catherine C. (1998). "Towards an ecology of hypertext annotation". In *Proceedings of HyperText 98*. New York: ACM. pp. 40-9.

Marshall, Catherine C. and A.J. Bernheim Brush. (2005). "Exploring the Relationship between Personal and Public Annotations". In *Proceedings for the JCDL '04 conference*. New York: ACM.

Pandit, Naresh R. (1996). "The Creation of Theory: A Recent Application of the Grounded Theory Method". In *The Qualitative Report*, Volume 2, Number 4, December, 1996. Online at http://www.nova.edu/ssss/QR/QR2-4/pandit.html.

Pliny 2007. Pliny project homepage. Online at http://pliny.cch.kcl.ac.uk/index.html

Shipman, Frank M., Catherine C. Marshall, and Thomas P. Moran (1995). "Finding and Using Implicit Structure in Human-Organized Spatial Layouts of Information". In *Proceedings of CHI 95*, Denver, Colorado, May 7- 11, 1995, pp. 346-353

The Harvester of Iconclass Metadata: a web service for subject classification and subject retrieval in cultural heritage information systems

Hans Brandhorst
Arkyves
jpjbrand@xs4all.nl

Etienne Posthumus

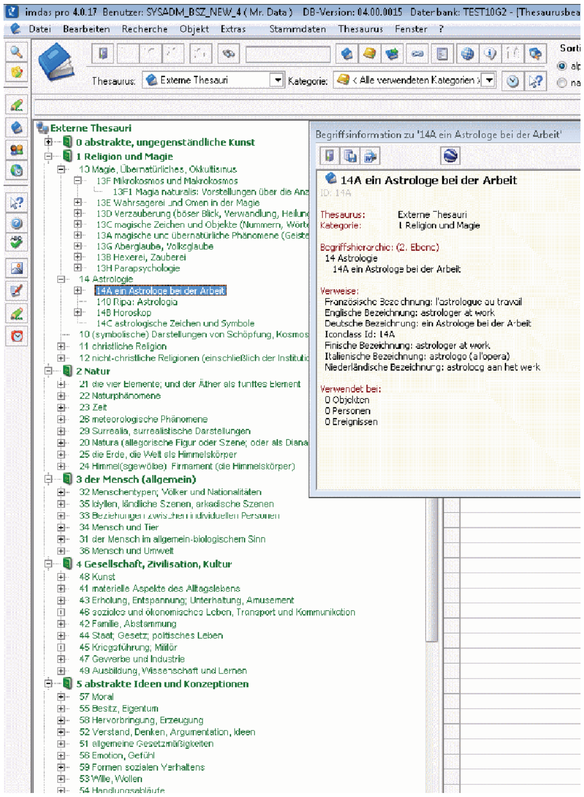
Summary

The importance of controlled vocabularies for the cataloguing of literary and visual sources by museums, libraries, and archives can hardly be exaggerated. Important organizations like the Library of Congress, the Getty Foundation, and OCLC have been promoting standardization by making thesauri and classification systems available online. However, complying to such a standard often means transferring descriptors from the online vocabulary system to another application, thereby isolating them from their context. While this is annoying during the production phase, it is fatal at retrieval time, because an end user should have access to the context of a descriptor while querying a catalogue. This poster demonstrates a web service for Iconclass, a multilingual subject classification system for cultural content that is used from Finland to Italy and from Germany to the US, which solves this problem. The service makes the Iconclass system available as an '*add on*' to database management systems and online electronic catalogues. The Iconclass metadata harvester uses OAI-PMH to gather Iconclass codes, supporting special retrieval browsers within local websites, while creating a single access point for thematic searches across multiple online databases.

1 The authority paradox

When using an online classification, a thesaurus or some other form of controlled vocabulary for cataloguing, we are sooner or later confronted by what could be labelled the *authority paradox*. The core of this paradox is that using a terminology authority in essence means *selecting specific terms* from the authority system and *copying* them to records of a local catalogue. However, the instruments that are available to cataloguers to select the most appropriate descriptors from the authority system, are *not* automatically available to the end user of a catalogue.

By definition vocabularies like classifications or thesauri —or, perhaps more fashionably, *ontologies*—are systems with a *structure*. This structure may be more or less complex, but at the very least there will be some basic inner organization. An example of such an organizational structure is that of the *broader and narrower* terms of a hierarchy. Simple hierarchical subordination is only one of various techniques to help a user find descriptors. Keywords may be added to help the user find the most appropriate concept; cross references may connect related concepts; redirects may point from non-preferred to preferred terms; scope notes explain the intention of a term, while links connect concepts across the languages of a multilingual system... these are just a few examples of additional, often quite sophisticated instruments to help one make the most of the authority system. No matter how simple or complex the structure of the authority system is, its default use will often be limited to the transfer of the standardized term for an artist, a placename, a profession, or an iconographical feature from the system to the catalogue. It would obviously cause an enormous amount of redundancy if one were also to copy all of the parent terms in a hierarchy, let alone if one were to transfer all the terms to which a chosen descriptor may be related—if that were at all possible.


imdas pro 4.0.17 Benutzer: SYSADM_BSZ_NEW_4 (Mr. Data) DB-Version: 04.00.0015 Daterbank: TEST10G2 - [Thesaurusbea
Datei Bearbeiten Recherche Objekt Extras Stammdaten Thesaurus Fenster ?
Sorti
alp
na

Thesaurus: Externe Thesauri
Kategorie: < Alle verwendeten Kategorien >

```
Externe Thesauri
  0 abstrakte, ungegenständliche Kunst
  1 Religion und Magie
    13 Magie, Übernatürliches, Okkultismus
      13F Mikrokosmos und Makrokosmos
        13F1 Magia naturalis: Vorstellungen über die Anz
      13E Wahrsagerei und Omen in der Magie
      13D Verzauberung (böser Blick, Verwandlung, Heilung
      13C magische Zeichen und Objekte (Nummern, Wörte
      13A magische und übernatürliche Phänomene (Geiste
      13G Aberglaube, Volksglaube
      13B Hexerei, Zauberei
      13H Parapsychologie
    14 Astrologie
      14A ein Astrologe bei der Arbeit
      110 Ripa: Astrologia
      14B Horoskop
      14C astrologische Zeichen und Symbole
    10 (symbolische) Darstellungen von Schöpfung, Kosmos
    11 christliche Religion
    12 nicht-christliche Religionen (einschließlich der Institutio
  2 Natur
    21 die vier Elemente; und der Äther als fünftes Element
    22 Naturphänomene
    23 Zeit
    26 meteorologische Phänomene
    29 Surrealia, surrealistische Darstellungen
    20 Natura (allegorische Figur oder Szene; oder als Diana
    25 die Erde, die Welt als Himmelskörper
    24 Himmel(sgewölbe) Firmament (die Himmelskörper)
  3 der Mensch (allgemein)
    32 Menschentypen; Völker und Nationalitäten
    35 Idyllen, ländliche Szenen, arkadische Szenen
    33 Beziehungen zwischen individuellen Personen
    34 Mensch und Tier
    31 der Mensch im allgemein-biologischem Sinn
    36 Mensch und Umwelt
  4 Gesellschaft, Zivilisation, Kultur
    48 Kunst
    41 materielle Aspekte des Alltagslebens
    43 Erholung, Entspannung; Unterhaltung, Amusement
    46 soziales und ökonomisches Leben, Transport und Kommunikation
    42 Familie, Abstammung
    44 Staat; Gesetz; politisches Leben
    45 Kriegsführung; Militär
    47 Gewerbe und Industrie
    49 Ausbildung, Wissenschaft und Lernen
  5 abstrakte Ideen und Konzeptionen
    57 Moral
    55 Besitz, Eigentum
    58 Hervorbringung, Erzeugung
    52 Verstand, Denken, Argumentation, Ideen
    51 allgemeine Gesetzmäßigkeiten
    56 Emotion, Gefühl
    59 Formen sozialen Verhaltens
    53 Wille, Wollen
    54 Handlungsabläufe
```

Begriffsinformation zu '14A ein Astrologe bei der Arbeit'

14A ein Astrologe bei der Arbeit
ID: 14A

Thesaurus: Externe Thesauri
Kategorie: 1 Religion und Magie

Begriffshierarchie: (2. Ebene)
14 Astrologie
14A ein Astrologe bei der Arbeit

Verweise:
Französische Bezeichnung: l'astrologue au travail
Englische Bezeichnung: astrologer at work
Deutsche Bezeichnung: ein Astrologe bei der Arbeit
Iconclass Id: 14A
Finische Bezeichnung: astrologer at work
Italienische Bezeichnung: astrologo (a l'opera)
Niederländische Bezeichnung: astroloog aan het werk

Verwendet bei:
0 Objekten
0 Personen
0 Ereignissen

Based on the simple fact that it copies its descriptors from '*authority system X, Y or Z*', a catalogue may claim to be compliant with that authority. However, this procedure of transferring terms, isolating them from the structure that embeds them, clashes with the purpose of complying with an authority system. Or, at the very least, the resulting catalogue will fail to offer to its end users the features that turn authority systems into actual 'systems'.

2 SKOS version of Iconclass as an external thesaurus add on

How the web service of Iconclass—in a SKOS version —functions as an *add on* may be efficiently demonstrated with the help of the preceding illustration, a screenshot of the *Imdas* database management system which offers Iconclass as an *external thesaurus* to its users.

Using a simple i-Frame and a local style sheet, Iconclass is shown—here in its German form—as a tree of concepts inside the *Imdas* application. Any browse action refreshes the content of the tree by triggering a request for updated information to the Iconclass server.

3 Restoring the link for retrieval: the Harvester for Iconclass Metadata (HIM)

The previous illustration suggests how terms may be transferred from the external authority system to the local database at *production time*. It also suggests that the transfer of the alphanumeric notation that accompanies every concept and assigns it a unique place in the hierarchy, suffices to link a database record to a concept and its complete context in the vocabulary system. It does *not* illustrate how the link between the individual terms that are copied to the application database and the Iconclass system is restored at *retrieval time*.

Before we look at how the *Harvester for Iconclass Metadata (HIM)* service restores this link, we should list some rather obvious assumptions:

A. The catalogue is available on the internet or at least on an intranet that is linked to the internet.

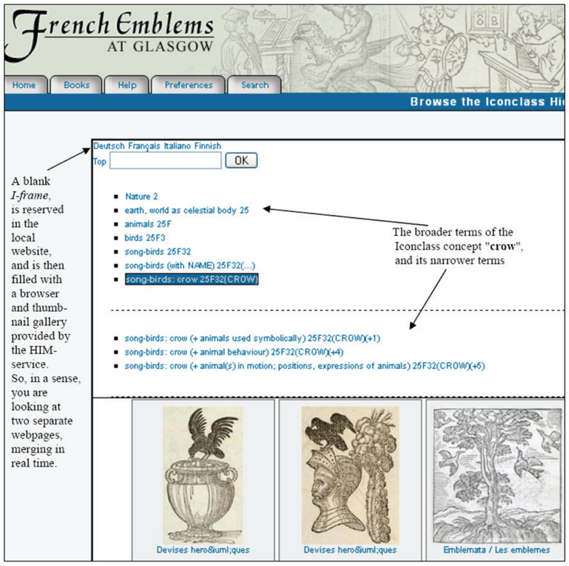
B. All items in the catalogue are identifiable with the help of some unique property, e.g. an inventory number.

C. This unique property can be used to retrieve an item from the catalogue.

Needless to say that the assumption about the Iconclass system is that this is a web service, permanently available on the internet. It may be consulted by human cataloguers, but it is also available for information exchange between computers.

Back to the question of how to restore the link between the copied terms in an application database and the Iconclass system's server. Actually, the answer is quite simple. Although it is theoretically possible to enrich a catalogue by absorbing major parts—or even the whole—of the Iconclass system, by far the easier strategy is to reverse the procedure, limit the information stored in the local database to Iconclass notations and then enrich the Iconclass system with information about the catalogue. What makes it so easy to export information about a catalogue that uses Iconclass for its subject access, is the simple fact that Iconclass is a classification system. Therefore, every concept in the system, with all of its links to other concepts and its translations, corresponds to a single code, or 'notation'. Like barcodes or ISBN-numbers these notations are thus very concise containers of information. These concise containers can easily be harvested using the *Open Archives Initiative's Protocol for Metadata Harvesting* (OAI-PMH) or customized variants thereof.

Although the Iconclass codes and the unique identifier of the object (e.g. a catalogue item) to which they have been assigned can be supplemented with other types of metadata, the Iconclass codes and the identifier are the most essential requirements for the harvesting service to work.



A blank *I-frame*, is reserved in the local website, and is then filled with a browser and thumbnail gallery provided by the HIM-service. So, in a sense, you are looking at two separate webpages, merging in real time.

- Nature 2
- earth, world as celestial body 25
- animals 25F
- birds 25F3
- song-birds 25F32
- song-birds (with NAME) 25F32(...)
- song-birds: crow 25F32(CROW)

The broader terms of the Iconclass concept "**crow**", and its narrower terms

- song-birds: crow (+ animals used symbolically) 25F32(CROW)(+1)
- song-birds: crow (+ animal behaviour) 25F32(CROW)(+4)
- song-birds: crow (+ animal(s) in motion; positions, expressions of animals) 25F32(CROW)(+5)

Devises hero&iuml;ques

Devises hero&iuml;ques

Emblemata / Les emblemes

The illustration above summarizes in a single picture the essential elements of the service. What you see there is the first row of a thumbnail gallery incorporated in the *French Emblems* website at Glasgow University (http://www.emblems.arts.gla.ac.uk/french/search.php). Above the thumbnails you see the concept "*song-birds: crow 25F32(CROW)*" highlighted. Its broader terms are listed above it. Below it, its narrower terms that were actually used for this catalogue are also listed. Whenever a concept is clicked, the corresponding Iconclass notation is sent to the central Iconclass server. The server then returns the object identifiers to the local database. These identifiers are subsequently used to retrieve the corresponding objects from the local database. In the small box at the top simple and complex keywords searches may be entered in the various languages of Iconclass, i.e. English, German, Italian, or French (partial translations exist in Finnish and Dutch).

By providing Iconclass through a web service, all users have access to the same, i.e. the present version of the system. Editorial changes are instantly available. If a local database merely stores codes, its users have access to the various languages of the system, and the full context of every concept—an efficient way to overcome the *authority paradox*.

The aggregator website *www.arkyves.org*—a single access to over 150,000 objects indexed with Iconclass—will be shown as part of the poster, in addition to the Iconclass web service.

Iconclass computing:
Arkyves
W.G. Plein 124
1054 SG Amsterdam
The Netherlands
tel +31 20 616 1039
e-mail: info@arkyves.org
website: http://www.arkyves.org

Iconclass content:
Rijksbureau Kunsthistorische Documentatie
PO Box 90418
2509 LK The Hague
The Netherlands
tel +31 70 33 39 777
fax +31 70 33 37 789
e-mail: iconclass@rkd.nl
website: http://www.rkd.nl

Appendix—reaction to reviewers' comments

Review 1: A somewhat outdated description of the basic idea of the Harvester of Iconclass Metadata can be found at: http://mnemosyne.org/IIHIM/overview.rst.html

Although some details and names have changed, the underlying principle is unchanged: classification codes and a unique identifier for the item to which the codes (desciptors) have been assigned are harvested and stored in a central database. The codes are parsed and interpreted. All of its implicit properties (textual definitions and their translations, keywords, hierarchical links, cross references, etcetera) are then extracted from the Iconclass datafile and linked to the code. All search and browse actions at the client's website are then compared with the information stored in the central database and results are sent back to the client's system. This procedure is necessary for two reasons: A) technically Iconclass is a complex system and it would be very expensive—and redundant—to create search and browse software for each client system; and B) for a single catalogue item many descriptors (sometimes dozens of codes) may be used simultaneously. The textual information and other properties implicit in all codes that were assigned to a single item have to be made available to the end user simultaneously. The way to manage this efficiently is by storing all information in a central database.

Due to special features that were designed prior to the digital age, Iconclass is technically more complex than most classifications, so if the software works for Iconclass—which it does—it can in all likelihood cope with other classification systems as well. The add-on is a proprietary tool, but it is made available for free to any institution which is prepared to share (part of) its (meta)data.

Review 2: Since there is only one online version of the vocabulary all changes are made centrally. Most changes will be additions of more specific concepts, so they won't affect existing documents. At most older documents will not take full advantage of increasing specificity in later versions of Iconclass, unless they re-edit certain descriptions. Existing concepts have almost never been withdrawn or given new meaning in the 35 years of Iconclass' usage history.

There is no room here to expand on the idiosyncracies of Iconclass as a classification, but their presence is well documented in earlier literature (in particular in special issues of Visual Resources). The complexity of the code structure sets Iconclass rather apart as a classification. However, there are some parallels with biomedical searching techniques (Collexis, fingerprinting, Knowlets, WikiProfessional) that may be worth investigating. It goes without saying that, financially, these tools are in a very different league...

The ambition of HIM is not to be innovative per se, but to offer a cost-effective solution to a problem—"how to make the most of the use of Iconclass"—that would otherwise require expensive software and the reinvention of the wheel for every collection using the system. In the world of the Humanities that in itself may be seen as an innovation.

Modelling the Prosopography of the Royal Portuguese Court in the Sixteenth Century

Andreia Carvalho
King's College London
andreia.carvalho@kcl.ac.uk

The poster will describe the digital component of the research project developed in the course of my PhD in Digital Humanities at King's College London.

The project features a prosopographical study of the high officials of the Royal Court during the reign of John III, king of Portugal (1521-1557). The poster will demonstrate how a relational model was used for that purpose: the research aims at identifying a particular group of people, detecting their socio-economic characteristics, and uncovering their patterns of behaviour in the context of early modern court politics.

Despite being a small country, in the sixteenth century Portugal possessed a large royal court. I am interested in the human configuration of this structure; who are the men who govern the kingdom and surround the king? This question can be addressed through an analysis of their offices and of their social and economical background. A further analysis of the nature of the relationship between these actors and their (collective or individual) relationship with the king will also provide insights on the nature of influence and the negotiation of power in royal courts during the Renaissance.

The ultimate aim of the project is to publish an online version of the database, which will make it available to other researchers as well as the general public. However, this poster will describe my research rather than focus on the implementation of that database.

The use of relational databases for prosopographical purposes has been a recurrent practice in historical research since the 1980s. At the Centre for Computing in the Humanities there are several projects devoted to prosopography. Both the Prosopography of Anglo-Saxon England (PASE) and the Prosopography of the Byzantine World (PBW) have now become leading resources in the academic world and exemplars in this practice. Having emerged from the specific needs and constraints required by historians, the databases have now evolved in order to allow new avenues of enquiry that go beyond the initial aims of the researchers. Both projects have become became laboratories for the relational modelling of historical sources. This led to the development to what Bradley and Short called a 'highly structured approach to historical data', or 'new-style prosopography' (Bradley and Short 2005).

One of the major challenges of this project was to test, adapt and revise the relational model used in PASE and in PBW.

A prosopographical database is essentially comprised of three elements: *sources* containing '*factoids'* about *persons*. These elements should provide a guide to the modelling stage. The poster will focus on specific issues:

1. How to incorporate different types of sources? The research will include published sources, original (manuscript) archival material and edited manuscripts. The sources are of a wide range, mainly literary and administrative. This diversity challenges, for instance, notions of authorship. Who is the author of a royal letter: the king, the person who drafted the letter, or all the officials who authorized and signed the document?

2. How to deal with 'factoids' involving several actors/agents? In the relational databases developed at King's College London, a factoid is an 'assertion by a source about one or more persons' (Bradley and Short 2001). The database revises the solution adopted in PASE and PBW, where an 'event', although being a 'factoid', was effectively separated from the 'factoid' table. The model presented here reconfigures these tables and adopts a different approach. The solution was to introduce an intermediary table called 'PersonInFactoid', which combines 'factoids' and persons, associated with an authority table that specifies the role played by each person in a given 'factoid'.

3. How to organize the different types of 'factoids'? This is done by the use of authority lists that in fact can 'manage' the 'factoids'. The database retains the authority tables used in PASE while simplifying and at the same time tightening their structure.

4. Finally, can it be possible to design a database that allows users to search for things not envisaged when doing the initial research? In fact, either the developer or the future users might want to access further data. Although the diverse types of factoids reflect the specific research questions of the project they do not compromise the assertions conveyed by the sources – these are kept in the factoid table.

The process of analysis and database design produced a ‘lighter’ and modified version of the PASE database structure. The poster will show the different tables created and their relationships, highlighting the problems and constraints of modelling the sources to the digital model. The input process will be done using a source-driven approach, that is, the information in the database will be information found in the *corpora* of sources used.

References

Prosopography of Anglo-Saxon England (PASE): http://www.pase.ac.uk (accessed 14 November 2008).

Prosopography of the Byzantine World (PBW): http://www.pbw.kcl.ac.uk (accessed 14 November 2008).

Bradley, J. and Short H. (2001). Using Format Structures to create complex relationships: the prosopography of the Byzantine Empire – a case study, http://www. cch.kcl.ac.uk/legacy/staff/jdb/papers/pbe-leeds/body.html, accessed 14 November 2008.

Bradley, J. and Short H. (2005). Texts into Databases: The Evolving Field of New-style Prosopography, *Literary and Linguist Computing* 20: 3-24.

Inventing the Future of AI for Games: Lessons from EMPath

Sherol Chen
University of California, Santa Cruz
sherol@soe.ucsc.edu

The space of Narrativity has had the benefit of being explored for centuries. Technology, however, introduces a new factor to pioneer by expanding the possibilities of both the experiencing and the telling of a story. Advancements of AI provide a multitude of techniques for the process of authoring and the actual authoring of stories in games. This process, however, is not so straight forward, and has proved to be extremely challenging in practice. Narratives in games, although sharing similar narrative qualities of its medium predecessors, deliver a highly interactive experience, making games a matter of new media and new analysis. The purpose of this talk is to demonstrate and give an idea of what the contrast is between the theories and models created in academia and the practice of industry in using technology to express compelling and believable experiences.

Introduction

With emphasis on the advancements of believability in areas of sound and scene, the advancements in the believability or complexity toward intelligent interactions within commercial video games are comparably lacking. Certain games aim to leave the story up to the user’s imagination, while others reduce user agency to create an artistically inspired story, and still, others are focused on procedurally creating a story that has both high user agency and dramatic significance. Arguments that support story games demonstrating high levels of intelligence claim that with the incorporation of Drama Managing, Multi-Agent Systems, Reactive Planning, and other sub-areas of AI will come a new frontier in experiencing and telling stories reducing traditional conflicts (Mateas 2001). On the other hand, game developers in industry are resistant against these ideas due to the complications that come along with adding such methods. Efforts in commercial games have been more successful in making the most out of scripted stories, maximizing from sound and scene, and sacrificing certain qualities in order to strengthen others. In addition to surveying the areas that are currently researched in academia and discussing games developed through research, there will be a demonstration of the EMPath project. EMPath is a prototype sized, adventure game that utilizes the Declarative Optimization-Based Drama Manager (DODM). This prototype game, developed at UC Santa Cruz, is a real-time playable game that uses the DODM architecture and has been tested on over 100 users. The purpose of this demo is to exhibit a novel AI-based approach to interactive storytelling, as well as provide a concrete example of the challenges in the design process. Figure 1 is a screen shot of the dungeon map in the original EMPath game (Sullivan 2009).

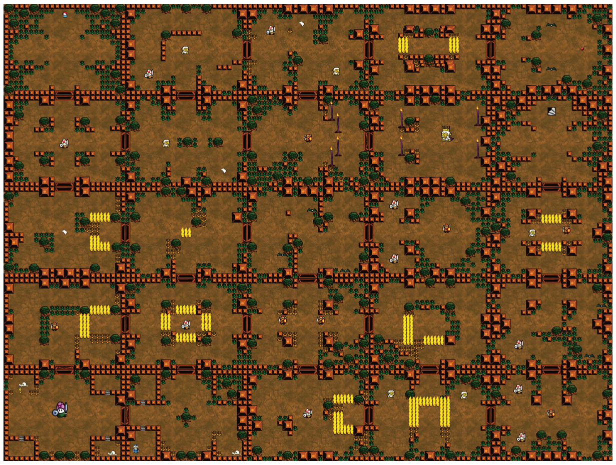


Figure 1. The 5x5 map world of EMPath.

Drama Management

A drama manager (DM) monitors an interactive experience, such as a computer game, and intervenes to shape the global experience so that it satisfies the author's expressive goals without decreasing a player's interactive agency. Most research work on drama management has proposed AI architectures and provided abstract evaluations of their effectiveness. A smaller body of work has evaluated the effect of drama management on player experience, but little attention has been paid to evaluating the authorial leverage provided by a drama management architecture: determining, for a given DM architecture, the additional non-linear story complexity a DM affords over traditional scripting methods. This poster will propose three criteria for evaluating the authorial leverage of a DM: 1) the script-and-trigger complexity of the DM story policy, 2) the degree of policy change given changes to story elements, and 3) the average story branching factor for DM policies vs. script-and-trigger policies for stories of equivalent quality. Figure 2 illustrates the decision-tree representation approach in evaluating authorial leverage. For preliminary studies in evaluating complexity of the drama manager, thousands of partial stories were generated and used to train and test decision trees using the J48 algorithm implemented in Weka, a machine-learning software package.1 Partial stories (the independent variable) are represented by a set of boolean flags indicating whether each plot point and DM action has happened thus far in this story, and, for each pair of plot points a and b, whether a preceded b if both happened (Chen 2009).

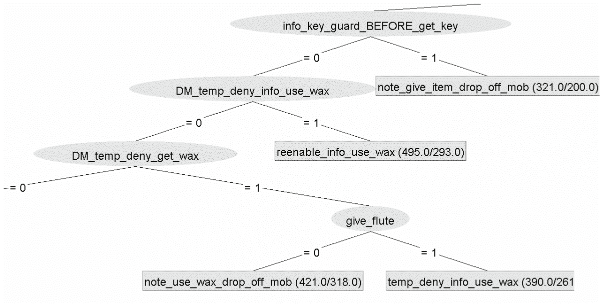

info_key_guard_BEFORE_get_key
= 0
= 1
DM_temp_deny_info_use_wax
note_give_item_drop_off_mob (321.0/200.0)
= 0
= 1
DM_temp_deny_get_wax
reenable_info_use_wax (495.0/293.0)
= 0
= 1
give_flute
= 0
= 1
note_use_wax_drop_off_mob (421.0/318.0)
temp_deny_info_use_wax (390.0/261

Figure 2. Zoomed-in view of a decision tree that has learned from the DM.

Lessons from EMPath

Gains for creating a more intelligently interactive story need to be proven through subject testing and other types of evaluation. In particular, there are two metrics that need to be established: one to show overall improved experience, and one to show the lightening of authorial burden. These approaches are demonstrated in previous experiments with EMPath; however, the prototype games that are feasible for research have difficulty demonstrating significant results due to the scale of these smaller games. Other challenges that arise from implementing and evaluating AI systems are as follows (Chen 2009):

- **Authoring:** Incorporating an AI system creates the burden of domain understanding, forcing an author to break traditional habits in authoring.

- **Evaluation Measures:** Story qualities must be mathematically represented in order to show improvement or to encourage better interactions.

- **Player Modeling:** AI systems often depend on predicting and anticipating user actions and motivations. The system would need to model human tendencies mathematically.

- **Simulation:** Experiments designed to run offline, according to the system's user model are needed to provide authorial leverage and sanity checks for story evaluation comparisons.

- **User Testing:** Users need to notice a difference in their overall experience when using the AI system versus not using the AI system. Experiments need to be carefully designed to show improvement in the user data.

- **Game Design:** The game must be designed to be able to demonstrate such differences. In general, games that are created for research purposes are often not expansive or large enough to show significant results.

- **Trail Blazing:** Finally, a great challenge in creating AI systems is that evaluation measures and user study approaches have not been rigorously tested for these purposes

Dimensions of Narrativity & Interactivity

Taking another look at Narratology, establishing a more complete understanding of narrative may help ameliorate some of the challenges in designing and testing AI systems. If there is a better understanding of the objectives that an AI system is aiming toward, then the space of story that it creates may be more easily evaluated. For instance, reducing the space of possible stories by fixing the ontological variations in an interactive story space reduces the variety of experiences, but results in a more focused output. With a fixed and linear diegetic universe, both the author's artistic vision is maintained and, as a result, the dramatic significance of the vision. Research can begin by finding ways of delegating types of discourse actions and performing them on a fixed set of plot points contingent on the actions of the user in trying to optimize user agency under those conditions.

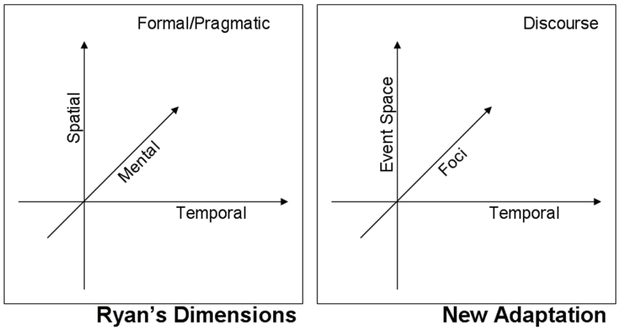


Figure 3. Adaptations of Ryan's 8 narrative dimensions.

By establishing narratives as a relationship among scalar properties, Marie-Laure Ryan's dimensions of narrativity gives a solid means to compare and analyze interactive narratives (Ryan 2006). For Ryan, her dimensions are more to establish conditions for the mark of the narrative. For the purpose of the discussion, a variation of Ryan's dimensions will be used analyze the experiential variations that result from interactivity in narratives. Figure 3 visually illustrates the analogous dimensions.

The new adaptation, instead of being a model for narratives, will serve as a model for interactive experiences. The four dimensions are: temporal (an axis to situate time), event space (an axis for delineating the occurrences in a story space), foci (the experiential views from intelligent existents or perspective), and discourse (an axis for determining the means of how a story is told). For further explanation of the relationship between these dimensions, it helps to "fix" or ignore one or more of the dimensions.

Conclusion

Overall, this presentation will deliver a broad understanding of the ways in which AI can integrate with interactive stories and create a diversity of experiences and outcomes. In contrast to commercial games, the process of interactive storytelling provides insights into the endeavors of universities and research institutions pioneering the area through advancements in AI. Additionally, there will be a live demo of the EMPath project in conjunction to a discussion of the difficulties and challenges in the counter-intuitive design process of a story that utilizes concurrent technologies of AI. Ultimately, this demo will be a case study towards gaining a deeper understanding of the challenges to be faced before what is possible in storytelling can be made practical through the intersections of the arts, sciences, industry, and academia.

Notes

[1]http://www.cs.waikato.ac.nz/ml/weka/

References

Chen, Nelson, Mateas (2009). Evaluating the Authorial Leverage of Drama Management. AAAI Spring Symposium, Interactive Narrative II.

Chen, Sullivan, Nelson, Wardrip-Fruin, Mateas (2009). Intelligent Interactive-Stories: Theory versus Practice. Game Developers Conference, San Francisco, CA.

Mateas (2001), A preliminary poetics for interactive drama and games. Digital Creativity, vol 12, number 3.

Ryan (2006). Narrative, Media, and Modes: Avatars of Story. University of Minnesota Press.

Sullivan, Chen, Mateas (2009). Abstraction to Reality: Integrating drama management into a playable game experience. AAAI Spring Symposium, Interactive Nar-

rative II.

Fine Rolls in Print and on the Web: Progress on a Reader Study

Arianna Ciula
King's College London
arianna.ciula@kcl.ac.uk

Tamara Lopez
King's College London
tamara.lopez@kcl.ac.uk

Faye Thompson
King's College London
faye.thompson12@hotmail.co.uk

This poster will report on progress made in the Reader Study related to the use of the Henry III Fine Rolls project resources. A poster on the research framework and phase one of the study was presented at the Digital Humanities conference in 2008. Since then (June 2008), the authors have continued work on later phases of the study, completing the document analyses, defining a research profile out of the sample data, and evaluating, developing and revising the methodological framework. This poster will report and comment on these findings.

Project Summary

The Henry III Fine Rolls (http://www.frh2.org.uk) is a collaborative project funded by the Arts and Humanities Research Council (UK) between the National Archives in the UK, the History and Centre for Computing in the Humanities departments at King's College London, the Department of History and American Studies at Canterbury Christ Church University. At the core of the project is the study of the medieval primary sources known as the Fine Rolls. Dating back to the 1170s, these documents written in Latin on membranes of parchment were issued by the royal Chancery to record agreements made with the king to pay a certain sum of money for specific benefits. Those that witness writs for the whole reign of Henry III (1216-1272) were never published properly and in their entirety before the Henry III Fine Rolls project took the initiative to do so.

The outcome of the project, currently in its second phase, is both a resource website and a set of printed volumes[1] containing the calendared edition (an English summary of the Latin records) of the Fine Rolls.

Reader Study

This dual editorial effort has raised questions about presentational formats pertinent to the two media and presented challenges for the historians as well as the digital humanities researchers involved in the project and the publisher. The process of negotiating and envisaging different 'material' solutions for the two types of published resources presented the authors with an opportunity to examine this parallel production process and to inquire about the social processes that influence the ways in which scholarship is embedded in published outputs, both in print and digital form.

Therefore, the reader study was conceived to:

- Reflect on the material actualisations of the Fine Rolls hybrid edition;[2]
- Evaluate the use of this hybrid edition to tackle old and new research questions;
- Establish the effectiveness of particular features of each medium in creating *bridges* between the print and web resources;
- Articulate general heuristics that can be used in the design of other hybrid digital humanities projects.

As summarised below, the poster presented in 2008 developed a framework for this exploratory study using established data collection methods from information seeking research, in particular those developed for the report *Scholarly Work in the Humanities and the Evolving Information Environment* (Brockman, Neumann, Palmer & Tidline, 2001).

This methodological approach is qualitative, striving not to achieve statistical significance but rather to develop a nuanced picture of the work performed by scholars using this and like material. The decision to use qualitative methods has forced the authors to limit the number of participants to a small group, collecting information by email and face-to-face interviews and through content analysis of research material. The participant sample throughout draws from members of the Fine Rolls research team who have written materials using the edition, but also on a small sample of researchers from the wider scholarly community who perform text-based archival research with documentary primary sources that date back to the late Middle Ages.

Analysis of the materiality of the fine rolls edition in 2008 was coupled with an initial examination of distributed questionnaires and document analysis of material written by study participants on the use of this or similar editions. Emerging trends from this exercise suggested that scholars do posses a high familiarity with primary sources through continuous direct access. In addition, the authors found that participants tend to cite a core set of sources over and over again. These key sources mainly consist of facsimiles, editions or reference works that give access to a substantial corpus of primary sources in various forms.

This poster will report more specifically on how scholars from this community perceive and overcome obstacles related to access or use of materials in performing research, and share general attitudes about what constitutes a successful research process (Dervin, 1992; 1998). In addition, it will identify and report on information patterns for the community, defining the types and formats of sources that are used, including insights into use of access resources (Brockman et al., 2001; Palmer, 2005). Findings from this phase will confirm whether or not the research process of participants conforms to Palmer's mode of access for humanities scholars (2005).

In addition, this poster will isolate the bridging tactics (Wilson, 1999) employed by scholars when moving between digital and print materials, in order to more fully develop a lexicon of connective structures for hybrid editions. Finally, a preliminary testing plan for evaluating the design of the Fine Rolls will be presented.

Bibliography

Brockman, W., L. Neumann, C. L. Palmer, T. J. Tidline. (2001) *Scholarly Work in the Humanities and the Evolving Information Environment*. Digital Library Federation Council on Library and Information Resources, December 2001, Council on Library and Information Resources.

Ciula, A. and Lopez T. "Reflecting on a Dual Publication: Henry III Fine Rolls Print and Web". *Literary and Linguistic Computing* (forthcoming).

Dervin, B. (1992) "From the mind's eye of the user: the sense-making qualitative-quantitative methodology". In Glazier, J.D. and Powell, R.R. (eds.) *Qualitative Research in Information Management.Libraries Unlimited*, 1992. pp. 61-84. Pre-print accessed 14 November, 2008 at: http://www.ideals.uiuc.edu/html/2142/2281/Dervin1992a.htm.

Dervin, B. (1998) "Sense-making theory and practice: an overview of user interests in knowledge seeking and use". *Journal of Knowledge Management*. Vol. 2 No. 2, December 1998. pp. 36-46.

Dryburgh, P., Hartland, B. eds., Ciula, A., Vieira J. M. tech. eds. (2007) *Calendar of the Fine Rolls of the Reign of Henry III preserved in The National Archives. Volume 1: 1216–1224***.** Woodbridge: Boydell & Brewer.

Dryburgh, P., Hartland, B. eds., Ciula, A., Vieira J. M. tech. eds. (2008) *Calendar of the Fine Rolls of the Reign of Henry III preserved in The National Archives. Volume II: 1224–1234***.** Woodbridge: Boydell & Brewer.

Finkelstein, D. and McCleery, A. (2002) *The book history reader***.** London: Routledge.

Palmer, C. L. (2004) "Thematic Research Collections". In Schreibman, S., R. Siemens and J. Unsworth (eds.) *A Companion to Digital Humanities*. Blackwell Companions to Literature and Culture. Oxford: Blackwell, 2004. pp. 349-365.

Palmer, C. L. (2005) "Scholarly work and the shaping of digital access". *Journal of the American Society for Information Science and Technology* Vol. 56 No 11, 2005. pp. 1140-53.

Wilson, T. (1999) Models in information behaviour research. Journal of Documentation Vol. 55 No 3, 1999. pp. 249-270.

Notes

[1] The first two volumes (Dryburgh et al., 2007; 2008) have already been printed in collaboration with the publisher Boydell & Brewer, and a third is under production: another five volumes will be published to complete the edition.

[2] That is, extant both in a physical and virtual environment. For a fuller treatment of this, see Ciula and Lopez (forthcoming).

Digital Tools from the Center for History and New Media: Present and Future

Dan Cohen
Center for History and New Media
dan@dancohen.org

Tom Scheinfeldt
Center for History and New Media
tom@foundhistory.org

Jeremy Boggs
Center for History and New Media
jboggs@gmu.edu

Dave Lester
Center for History and New Media
dave@omeka.org

The need to "develop and maintain open standards and robust tools" was one of eight key recommendations in *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. It has become increasingly apparent that if digital humanists are to have the right tools for their work, they will have to build some of them themselves. Indeed, many humanities scholars and social scientists have already built valuable digital tools over the years, including software and systems for text processing, markup, visualization, metadata generation, cataloging, GIS, and a number of other humanities-related tasks.

In the past seven years, the Center for History and New Media at George Mason University has expanded its focus from creating web resources for educational, scholarly, and general audiences to include several major open source software projects. Making this transition has involved a considerable challenges, but also some significant opportunities. In this panel the directors and developers of these digital tools will speak about the challenges of creating software for scholarship, research, web publishing, and course management; discuss where their tools are today and how others can get involved in their production; and talk about future plans and trends. Three projects will be highlighted:

Zotero

For research management, CHNM has created Zotero (http://zotero.org), an easy-to-use yet powerful research tool that helps users gather, organize, and analyze sources (citations, full texts, web pages, images, and other objects), and share the results of their research in a variety of ways. An extension to the popular open-source web browser Firefox, Zotero includes the best parts of older reference manager software (like EndNote)—the ability to store author, title, and publication fields and to export that information as formatted references—and the best parts of modern software and web applications (like iTunes and del.icio.us), such as the ability to interact, tag, and search in advanced ways. Zotero integrates tightly with online resources; it can sense when users are viewing a book, article, or other object on the web, and—on many major research and library sites—find and automatically save the full reference information for the item in the correct fields. Since it lives in the web browser, it can effortlessly transmit information to, and receive information from, other web services and applications; since it runs on one's personal computer, it can also communicate with software running there (such as Microsoft Word). And it can be used offline as well (e.g., on a plane, in an archive without WiFi).

Zotero has surpassed the milestone of over a million users and more than a hundred colleges and universities now actively recommend Zotero to their students, faculty, and staff. The Zotero project has received recognition in PC Magazine's "Best Free Software" issue. Through the open source community, Zotero has been translated into 40 languages, ranging from Arabic to Vietnamese.

Omeka

For web publishing, CHNM has created Omeka (http://omeka.org), a free and open source collections based web-based publishing platform for scholars, librarians, archivists, museum professionals, educators, and cultural enthusiasts. CHNM's new open source platform for publishing collections and exhibitions online hit a major milestone with the release of the 0.10 Beta release. Designed for cultural institutions, enthusiasts, and educators, Omeka is easy to install and modify and facilitates community-building around collections and exhibits. It is designed with non-IT specialists in mind, allowing users to focus on content rather than programming. Its "five-minute setup" makes launching an online exhibition as easy as launching a blog. Omeka is designed with non-IT specialists in mind, allowing users to focus on content and interpretation rather than programming. It brings Web 2.0 technologies and approaches to academic and cultural websites to foster user interaction and participation. It makes top-shelf design easy with a simple and flexible template system. Its robust open-source developer and user communities underwrite Omeka's stability and sustainability. Until now, scholars and cultural heritage professionals looking to publish collections-based research and online exhibitions required either extensive technical skills or considerable funding for outside vendors. By making standards based, serious online publishing easy, Omeka puts the power and reach of the web in the hands of academics and cultural professionals themselves.

ScholarPress

As educators increasingly use blogs in their classrooms, CHNM has explored ways to provide open-source tools for educational blogging. Currently an unfunded, staff-generated project, ScholarPress (http://scholarpress.net) is a development hub for educational and scholarly plugins for the WordPress blogging platform. ScholarPress currently offers two plugins: Courseware, which enables teachers to manage a class with a WordPress blog with a schedule, bibliography, assignments, and other course information, and WPBook, which works with the Facebook Development platform to create an (addable by users within the site) using a Wordpress blog. Thus, someone using both Courseware and WPBook can create a WordPress blog, add a course schedule, reading list, assignments, and annoucements, and share that information with students through the blog and through a Facebook application that students add. ScholarPress plans future plugins for grading, research, bibliography management, and is working to make plugins compatible with multi-user versions of WordPress, so services like Edublogs can add the plugins and make them available to educations using their services.

"Digital Tools from the Center for History and New Media: Present and Future" will provide audience members with an introduction to each of these tools, the challenges involved in building each, the similarities and differences in building tools for different purposes and audiences, and some lessons learned in making the transition from web development to software development. The format for the session will feature short, 15 minute presentations by four key members of the CHNM tool development team. CHNM director Dan Cohen will discuss Zotero, Managing director Tom Scheinfeldt and Omeka Developer Dave Lester will discuss Omeka, and Creative Lead Jeremy Boggs will discuss ScholarPress.

These short project-focused presentation will be followed by a 15 panel discussion among the presenters in which they will discuss the similarities and differences between the three projects and some of the more general aspects of software development. The balance of the remaining time will be left to audience questions and answers. All four participants live in the Washington D.C. area and will be available during the week of DH09. CVs for each are attached.

Access versus Ownership

Navigating the Tension between Mass Digitization of Archival Materials and Intellectual Property Rights

Maggie Dickson
University of North Carolina, Chapel Hill
mdickson@email.unc.edu

Amy Johnson
University of North Carolina, Chapel Hill
johnsoal@email.unc.edu

Natasha Smith
University of North Carolina, Chapel Hill
nsmith@email.unc.edu

Lynn Holdzkom
University of North Carolina, Chapel Hill
uholro@email.unc.edu

Stephanie Adamson-Williams
University of North Carolina, Chapel Hill
sadamson@email.unc.edu

The advent of digital technologies has presented archivists with opportunities to provide unprecedented, and, increasingly, expected, online access to collections to their patrons. Some archival repositories are now exploring, and in some cases, such as the Archives of American Art Collections Online and the Library of Congress's American Memory Project, engaging in, the mass digitization and Web presentation of their collections, which allows patrons to perform much of the same research from their homes or offices at any time of day, for which they would traditionally have had to travel great distances and been subject to limited hours of operation.[1]

In 2007 the Carolina Digital Library and Archives (CDLA) and the Southern Historical Collection (SHC) began work on a pilot project to explore the most effective methods of mass digitization and gather data on which a full-scale program could be founded. A 2-year, $300,000 grant from the Watson-Brown Foundation of Thomson, GA funded the project. The correspondence series of the Thomas E. Watson Papers, housed in the SHC, was chosen as the subject of the pilot project. Thomas E. Watson, of Thomson, Ga., was a prominent Populist politician, author, and lawyer, and his correspondence series consists of 8 linear feet of letters and related material written by Watson and his family, friends, and political and business colleagues. The date range of the correspondence series is 1873-1986, with the bulk of the letters dating between the 1880s and 1920s.

One of the challenges facing this project is navigating between providing comprehensive online access while at the same time respecting the intellectual property rights of any copyright holders who may be represented in the collection. Unpublished manuscript materials, such as those found in the Thomas E. Watson Papers, are protected by copyright for 70 years plus the life of the author. For us, that means that any letters written by a correspondent who died prior to 1939 (2009 being the projected publication date for our digital collection) are fair game under general copyright rules. However, any letters written by a correspondent who died after 1939 are potentially still in copyright.

So, if we were not to claim any exemptions to copyright statutes, and we wanted to publish the entire correspondence series on the Web under a strict interpretation of copyright law, we would need to identify all of the authors of the letters in the series, determine the date of their deaths, locate their descendants if their death dates were after 1939, contact those descendants, and request and then obtain permission to use their deceased family members' letters.

Even though we suspected this effort had little chance of success, we determined that because our project is not only about digitizing the Thomas E. Watson Papers, but also serves as a pilot for a much larger effort aimed at digitizing the entire SHC, we would attempt to do intense copyright research on the materials in the series to investigate thoroughly this aspect of digitizing archival materials. Prior to beginning copyright research, we had already gathered basic metadata (correspondent and recipient names, places from which letters were written, and dates, for example) from all of the 8400+ documents. It took us approximately 91 hours to go through the 8 linear feet of materials in the correspondence series to compile this data. From the information we gathered, we were able to condense and regularize the correspondent list down to approximately 3300 names.

Using a variety of sources, including Wikipedia, the Social Security Death Index, Ancestry.com, and print references, we attempted to positively identify the 3300 names. What resulted was a list of 3280 confirmed and questionable identifications, and 24 unknowns that were simply impossible to identify. We were able to locate birth or death dates for 1709 of the identified correspondents, while for 1571 no dates were available via the consulted sources. Of the correspondents for whom we located dates, 1101 died after 1939, while 608 died during or before 1939.

Of the positively identified individuals who died after 1939, we found that just over 50 had dedicated manuscript collections (or materials in the manuscript collections of other individuals) deposited in repositories. In these cases, we contacted the repositories, asking for their latest acquisition information, in the hopes that it might lead us to descendents of the correspondents from whom we could request permission to digitize their relative's materials. In most cases, no information was available, and when it was, it tended to be outdated, often well over 20 years old. We were able to obtain current, dependable contact information for the descendents of only two of the more prominent correspondents – Upton Sinclair and Hamlin Garland, both of whom are well-known writers with established literary estates.

The fact that it was so difficult to obtain contact information for the descendents of people who are prominent enough to have dedicated manuscript collections indicates that locating the descendents of the bulk of the correspondents would be daunting and in most cases impossible.

Extrapolating from our experience with the Watson correspondence, we believe that an effort on the scale anticipated in digitizing the entire SHC would be stymied by trying to do in-depth exploration of copyright status and attempting to obtain permission to digitize unpublished archival materials that are under copyright. If we hope to make mass digitization an integrated part of processing archival materials, it is simply untenable for us to consider doing this type of research to determine and obtain copyright.

This does not mean that we should avoid digitizing materials which may still be in copyright, but it does mean that we need to find a different way to approach copyright law to accommodate our needs.

Relying on Fair Use

While copyright law was intended to protect creative expression, it was at the same time not intended to be an impediment to further creative expression. Because of that, there are limitations to exclusive rights and remedies in sections 107 to 122 of the Copyright Act that allow for the use of copyrighted materials under some circumstances. As the current copyright law was written in 1976, however, its authors did not anticipate the ways in which digital technologies would change the potential uses of copyrighted materials, and we must interpret these limitations and remedies to determine which might best apply to the mass digitization of archival materials.

A close examination of the possible limitations and remedies available to us in the law as it stands indicated that the most reasonable option for us is to use the fair use provision. Section 107 of the Copyright Act – Limitations on Exclusive Rights: Fair Use, states that 'use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching, scholarship, or research, is not an infringement of copyright.'[2] The Supreme Court, in its ruling in favor of the defendant in *Stuart v Abend*, stated that: 'fair use … permits courts to avoid rigid application of the copyright statute when, on occasion, it would stifle the very creativity which that law is designed to foster.'[3]

Weighing our project against the four fair use factors[4] and taking into account the existing case law (of which very little applies to archives and special collections) we have developed an argument which we feel allows us to legally publish our digitized manuscript collections online. Unfortunately, the only way to know with certainty that a use is considered a fair one is to have it resolved in a federal court. The thought of such a court battle constitutes a worst-case scenario for us, but given the precedents already set by the courts, we are unlikely to become involved in such a situation.

Given these circumstances, it is reasonable for us to continue to serve our patrons in the most effective ways possible by accepting this risk and forging ahead with mass digitization. In order to maintain the level of service researchers are increasingly coming to expect, it is imperative that archives and special collections forge ahead with mass digitization without fear of recrimination.

References

Copyright Act of 1976, 17 U.S.C. § 107

Erway, R., and Schaffner, J. (2007). Shifting Gears: Gearing Up to Get Into the Flow. Report produced by OCLC Programs and Research. Published online at: http://www.oclc.org/programs/publications/reports/2007-02.pdf

Stuart v Abend, 495 U.S. 207, 236 (1990)

Notes

[1] Ricky Erway and Jennifer Schaffner's paper, "Shifting Gears: Gearing Up to Get Into the Flow," discusses the changing expectations of archival user communities, as well as the changing role of the archivist in the face of developing digital technology.

[2] Copyright Act of 1976, 17 U.S.C. § 107. Limitations on exclusive rights: Fair use

[3] *Stuart v Abend*, 495 U.S. 207, 236 (1990)

[4] They are, in brief: 1. the purpose and character of the use, including whether such use is of commercial nature or is for nonprofit educational purposes; 2. the nature of the work itself [whether it is a factual or creative work]; 3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole; 4. the effect of the use upon the potential market for or value of the copyrighted work.

Implementing Greek Morphology

Helma Dik
University of Chicago
helmadik@mac.com

Richard Whaling
University of Chicago
rwhaling@uchicago.edu

In this poster we discuss the nuts and bolts of our implementation of Greek morphology in a five-million word corpus, that of the Perseus Greek texts. Many disparate elements, and the efforts of many different people have come together in this project. Dik & Whaling (2008) describe how initial data was gathered from multiple sources; the current paper describes what went into the final product:

Disambiguated Greek texts: Early Greek Epic and the New Testament

The two sources of data from which we were going to bootstrap our project came with their own specific features: The two disambiguated corpora used different data specifications, which had to be compared and made uniform, and brought up to the standard that we wanted. However, it did bring us two large swaths of data of 350K words in total with which to seed our part-of-speech analyser, TreeTagger.

Morphological analysis from the Perseus project

The training data alone were not going to be adequate to produce a full lexicon for TreeTagger. We decided to supplement it with the output from Perseus's Morpheus tool (Crane 1991) of all possible parses for the full corpus. This greatly enhanced TreeTagger's accuracy on rare words not encountered in the training data, but also generated many redundancies and inconsistencies which made it hard for TreeTagger to build a proper decision tree. It was a continuing dilemma to those involved in the project whether we should allow more correct input to eventually weed out the many incorrect parses, or to remove incorrect parses from the input directly. Some effort was made to remove incorrect input: A 28 MB lexicon was reduced to less than 23 MB, but this represented only a portion of problematic entries.

TreeTagger

TreeTagger (Schmid 1995) is the proprietary software we used to assign part-of-speech tags (in effect, a full morphological disambiguation in a ten-slot morphological code plus a lemma). We trained TreeTagger in the first instance on the basis of the New Testament data, and added 40,000 words total in representative 1000 word samples from the rest of the corpus. New sets of samples were prepared with TreeTagger disambiguation, for which we used earlier disambiguated samples, plus our initial New Testament samples, as input.

Disambiguation (internal)

On the basis of earlier work on Czech and other languages, we decided that 40,000 words would be an adequate sample. This disregarded the fact that most research in natural language processing is actually done on more homogeneous texts than our samples of Greek literature, such as Reuters news items. Clearly, a more homogeneous input makes for higher accuracy within the source corpus, but we had no such luxury. An early indication was the high accuracy rate achieved by TreeTagger when trained and tested on Homeric Greek, which is a highly homogeneous, formulaic, subset of our texts. Perhaps the Homeric corpus is in fact the best parallel to Reuters and similar corpora in modern languages - at least the accuracy was comparable.

In more practical terms, undergraduate students of Greek were hired to 'pick the right parse' from among possible parses identified by TreeTagger. The disambiguation interface allowed the students to signal alternative parses or lemmas if none of the TreeTagger choices was accurate. Next, in an 'admin' layer, items about which there were disagreements among the students or about which comments were entered, were highlighted for review, so that the principal investigator could review these items especially, prior to feeding fully disambiguated texts back to TreeTagger.

Implementation

The centerpiece of our implementation is a SQLite database backend, containing the tokens and parses for the full corpus. It connects the three major components of the system:

- The original Perseus XML files, in which the tokens have been given unique ids as follows, keeping intact all previous markup:

  ```
  <w id=”276565”>ὦ</w>

  <w id=”276566”>ἄνδρες</w>

  <w id=”276567”>Ἀθηναῖοι</w
  ```

- TreeTagger, which accepts token sequences from the database and outputs parses and probability weights, which are stored in their own table.

- PhiloLogic, which serves as a highly efficient search and retrieval front end, by indexing the augmented XML files as well as the contents of the linked SQLite tables. PhiloLogic's highly optimized index architecture allows near-instantaneous results on complex inquiries such as 'any infinitive forms within 25 words of (dative singulars of) lemma X and string Y', which would be a challenge for typical relational database systems.

For a concrete example, in a standard PhiloLogic search box, entering 'lemma:μῆνις' will produce this word from the first line of the Iliad, as will a search for 'pos:*fa*', as will a search for the original string, 'μῆνιν'. Criteria can be combined as well, so that 'lemma:μῆνις;pos:*fa*' produces only feminine accusative forms of the particular lemma μῆνις.

We continue to explore the possibility of natural language searching as a substitute or alternative to this highly technical way of querying the corpus, and will demonstrate our progress on this front at the conference. The goal is to make it possible for users to type 'feminine accusative' as opposed to 'pos:*fa*', which will remain daunting to all but the most determined.

Conclusion

We are happy to have disambiguated a large corpus, making available for the first time a large, representative corpus of Classical Greek for morphological searching in addition to searching by string and by lemma — integrated into the existing reading and browsing environment for the texts. However, we are now also prepared to start crowd-sourcing the long tail of incorrect parses. Besides looking up the statistically most probable parse according to TreeTagger and other possible parses, users can 'vote' to correct TreeTagger's chosen parses. Once these votes have been inspected and accepted into the main database, future updates to the corpus will reflect both these local corrections and, over the full extent of the corpus, a more accurate TreeTagger. It is our hope that with the assistance of our users we will approach higher and higher levels of accuracy, making this tool ever more useful to scholars of Classical Greek.

References:

Crane, Gregory (1991). Generating and parsing classical Greek. In *Literary and Linguistic Computing*, 6(4): 243-245, 1991.

Dik, Helma and Richard Whaling (2008) - Bootstrapping Greek Morphology. Digital Humanities 2008.

Schmid, Helmut (1995) - Improvements in part-of-speech tagging with an application to German. In Proceedings of the ACL SIGDAT-Workshop. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf

Website URL: http://perseus.uchicago.edu

Synergies: An Overview and Progress Report

Michael Eberle-Sinatra
Université de Montréal, President *Synergies*
michael.eberle.sinatra@umontreal.ca

This poster will offer an overview of the CFI-funded project *Synergies: The Canadian Information Network for Research in the Social Sciences and Humanities*, and offer a progress report on the developments that took place in the first year of its four-year funding cycle, to be accompanied by a live demonstration of the alpha version of the web-based search interface.

Synergies is a four-year project intended to be a national distributed platform with a wide range of tools to support the creation, distribution, access and archiving of digital objects such as journal articles. It will enable the distribution and use of social sciences and humanities research, as well as to create a resource and platform for pure and applied research. In short, *Synergies* will be a research tool and a dissemination tool that will greatly enhance the potential and impact of Social Sciences and Humanities scholarship.

Canadian social sciences and humanities research published in Canadian journals and elsewhere, especially in English, is largely confined to print. The dynamics of print mean that this research is machine-opaque and hence invisible on the internet, where many students and scholars begin and sometimes end their background research. In bringing Canadian social sciences and humanities research to the internet, *Synergies* will not only bring that research into the mainstream of worldwide research discourse but also it will legitimize online publication in social sciences and humanities. The acceptance of this medium will open the manner in which knowledge can be represented. On one plane, researchers will be able to take advantage of an enriched media palette—color, image, sound, moving images, multimedia. On a second plane, researchers will be able to take advantage of interactivity. And on a third plane, those who query existing research will be able to broaden their vision by means of navigational interfaces, multilingual interrogation and automatic translation, metadata and intelligent search engines, and textual analysis. On still another plane scholars will be able to expand new areas of knowledge such as bibliometrics and technometrics, new media analysis, scholarly communicational analysis and publishing studies. This poster will introduce the main goals of the *Synergies* project and the impact it will have on the production and dissemination of Canadian research.

Scholarly research and communication are undergoing an evolutionary transformation. Research environments, scholarly communication, knowledge sharing and services are moving to the digital and becoming network-oriented. This evolution raises many new questions about models of knowledge sharing. Emerging research environments, data providers, publishers, and libraries will need to develop and deploy a wide variety of new resource models to address these new realities. These resource models will lower the barriers to access and exploitation of research and information resources, serve the needs of individuals and both general and specialized communities, and integrate new models of publication, annotation, communication and knowledge sharing.

Synergies will provide a needed infrastructure for the Social Sciences and Humanities Research Council (SSHRC) to follow through its in-principle commitment to open access and facilitate its implementation by extending the current venues and means for online publishing in Canada. With *Synergies* in place the funding of journals based on dissemination effectiveness rather than sales levels will become both feasible for journals and possible as an evaluative criterion for SSHRC funding. The Canadian Federation for the Humanities and Social Sciences, with a membership of over 30,000, has also recently adopted a position in favor of open access and indicate the role that *Synergies* can play.

Alongside a new web interface and tools for accessing information produced in Canada (with the alpha version to be demonstrated alongside the poster), *Synergies* will be a digital publishing platform for scholarly publications, with its first goal being to offer digital publishing services prepared to international standards with the lowest cost possible for the editorial production side. This project will thus work as a sustainable, open, e-publication infrastructure for the academy.

In sum, this poster will contain an overview of the project, and progress reports on its five regional components. *Synergies* is the result of a collaboration among five core universities which have been working together for several years. With each partner bringing its own expertise to the initiative, a genuine collaboration resulted in an infrastructure which was conceived from the start as truly scalable and extendable. Each regional node will integrate the input of current and future regional partners in the development of *Synergies*, thus continuing to extend its pan-Canadian dimension. For instance, the PKP project will be introduced within the broader context of scholarly communications. The Ontario region will be presented as a case study, with particular emphasis on project integration with Scholars Portal, a digital library. (The other three regions will also be included in this progress report to the Digital Humanities community.)

Fostering Cultural Literacies through Digital Scholarship: The Yaddo Archive Project and Yaddocast as Multimodal Humanities Projects

Richard L. Edwards
Indiana University
edwards7@gmail.com

Micki McGee
Fordham University
mmcgee@fordham.edu

Yaddo and the Yaddo Archive

Founded in 1900 by the financier and philanthropist Spencer Trask and his wife Katrina, Yaddo is an artists' community located on a 400-acre estate in Saratoga Springs, New York. Its mission is to nurture the creative process by providing an opportunity for artists to work without interruption in a supportive environment. Since its first official season in 1926, the artists' retreat has hosted more than 5,500 artists, writers, composers and other creatie workers including legendary figures such as Milton Avery, James Baldwin, Leonard Berstein, Truman Capote, Aaron Copland, Philip Guston, Patricia Highsmith, Langston Hughes, Ted Hughes, Alfred Kazin, Ulysses Kay, Jacob Lawrence, Carson McCullers, Sylvia Plath, Katherine Ane Porter, Mario Puzo, Clyfford Still, and Virgil Thomson.

In 1999, in celebration of the community's centennial, Yaddo and the New York Public Library entered into a unique partnership to ensure that Yaddo's archive—more than 550 boxes of rare letters, journals, guest applications, photographs, artworks, sound recordings, and other ephemera—would be preserved for posterity and available to scholars and the public. The Yaddo Records, now housed in the NYPL's Manuscript and Archives Division, constitute a unique resource for scholars in the humanities and of organizational development in that they include intimate letters between Yaddo's director and many of 20th century's most distinguished artists and nearly complete records of the organization's operations from 1900 to the present. In addition, Yaddo has kept detailed records of the arrivals and departures of its guests for nearly every year of operations, making charting the organization's contribution in building social relationships more viable than would otherwise be expected.

Multimodal Humanities and *Yaddo: Making American Culture* Exhibition

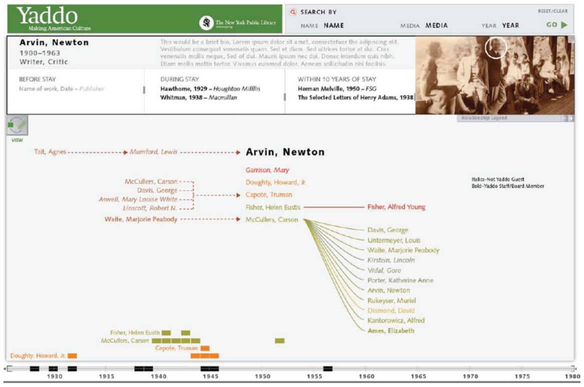
When Yaddo opened as an artist retreat in 1926, the *New York Times* hailed it as "a new and unique experiment, which has no parallel in the world of arts." That statement is still true today. Yaddo has not shed its experimental roots, and as part of its ongoing outreach efforts, it has dynamically begun to embrace 21st century technologies and techniques. And while the experience of the artists at Yaddo might not have dramatically changed in decades, the cultural outreach of Yaddo has been enchanced by digital scholarship projects in the multimodal humanities. David Theo Goldberg and Tara McPherson have advanced the concept of the "multimodal humanities." As McPherson states, the multimodal humanities "bring together databases, scholarly tools, networked writing and peer-to-peer commentary while also leveraging the potential of the visual and aural media that so dominate contemporary life." In the case of Yaddo, that media archive can encompass textual, visual, gestural, and aural works spanning the fields of literature, painting, composing, choreography, sculpture, film, video and mixed media.

As part of efforts to foster culture literacy around the opening of a new exhibition on the history of Yaddo at the New York Public Library in October 2008, the Yaddo Archive Project (a database project) and Yaddocast (an enhanced podcast series) were both created. In each project, database logics, participatory architectures, and interpretive spaces coalesce to tackle the complex question of Yaddo's impact on American culture. While there were different technical and production issues raised by these two projects, this paper will focus on the design of these projects as part of a larger epistemological question around how digital projects in the multimodal humanities mode can further our understanding of cultural literacies.

The Yaddo Archive Project

The Yaddo Archive Project (YAP) was not planned as the typical "exhibition website" that simply repurposes exhibition content with online presence. Rather, YAP aimed to use digital technologies to map key information in the Yaddo archive, from other records retained in Yaddo, and from secondary sources (artists' biographies, leters, and journals) to explore how Yaddo, as an artists' community, created cultural capital by fostering social capital.

The planned key components of YAP are an interactive online platform that 1) charts the network of relationships that made Yaddo a formidable force in 20th century American arts and letters, 2) map the relationships that were forged during Yaddo visits that later impacted American arts and letters, 3) charts art works (primarily published books, as those are most readily identified by date and accessed via online catalogs such as the Library of Congress and Worldcat) made before, during, and after Yaddo visits with an eye toward demonstrating effects of a Yaddo residency or residencies on subsequent artistic productivity, and 4) an access-protected entry point where scholars and archivists familiar with Yaddo and its artists can contribute new information on the relationships between members of this community (see Fig. 1 and Fig. 2).



Arvin, Newton
1900–1963
Writer, Critic

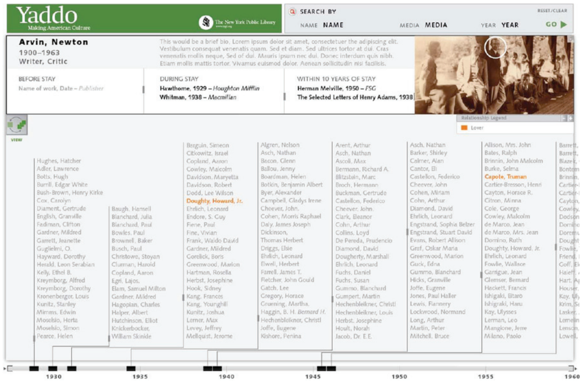
This would be a brief bio. Lorem ipsum dolor sit amet, consectetuer the adipiscing elit. Vestibulum consequat venenatis quam. Sed et diam. Sed ultrices tortor at dui. Cras venenatis mollis neque, Sed of dui. Mauris ipsum nec dui. Donec interdum quis nibh. Etiam mollis mattis tortor. Vivamus euismod dolor. Aenean sollicitudin nisi facilisis.

BEFORE STAY
Name of work, Date – *Publisher*

DURING STAY
Hawthorne, 1929 – *Houghton Mifflin*
Whitman, 1938 – *Macmillan*

WITHIN 10 YEARS OF STAY
Herman Melville, 1950 – *FSG*
The Selected Letters of Henry Adams, 1938

Fig. 1. Yaddo archive Project wireframe, 2 degrees of relationship.



Arvin, Newton
1900–1963
Writer, Critic

This would be a brief bio. Lorem ipsum dolor sit amet, consectetuer the adipiscing elit. Vestibulum consequat venenatis quam. Sed et diam. Sed ultrices tortor at dui. Cras venenatis mollis neque, Sed of dui. Mauris ipsum nec dui. Donec interdum quis nibh. Etiam mollis mattis tortor. Vivamus euismod dolor. Aenean sollicitudin nisi facilisis.

BEFORE STAY
Name of work, Date – *Publisher*

DURING STAY
Hawthorne, 1929 – *Houghton Mifflin*
Whitman, 1938 – *Macmillan*

WITHIN 10 YEARS OF STAY
Herman Melville, 1950 – *FSG*
The Selected Letters of Henry Adams, 1938

Hughes, Hatcher
Adler, Lawrence
Botts, Hugh
Burril, Edgar White
Bush-Brown, Henry Kirke
Cox, Carolyn
Diament, Gertrude
English, Granville
Fadiman, Clifton
Gardner, Mildred
Garrett, Jeanette
Guglielmi, O.
Hayward, Dorothy
Herald, Leon Serabian
Kelly, Ethel B.
Kreymborg, Alfred
Kreymborg, Dorothy
Kronenberger, Louis
Kunitz, Stanley
Mimms, Edwin
Moselsio, Herta
Moselsio, Simon
Pearce, Helen

Baugh, Hansell
Blanchard, Julia
Blanchard, Paul
Bowles, Paul
Brownell, Baker
Busch, Paul
Christowe, Stoyan
Clurman, Harold
Copland, Aaron
Egri, Lajos
Elam, Samuel Milton
Gardner, Mildred
Hagopian, Charles
Halper, Albert
Hutchinson, Elliot
Knickerbocker,
William Skinkle

Barguin, Simeon
Citkowitz, Israel
Copland, Aaron
Cowley, Malcolm
Davidson, Maryetta
Davidson, Robert
Dodd, Lee Wilson
Doughty, Howard, Jr.
Ehrlich, Leonard
Endore, S. Guy
Fiene, Paul
Fine, Vivian
Frank, Waldo David
Gardner, Mildred
Gorelick, Boris
Greenwood, Marion
Hartman, Rosella
Herbst, Josephine
Hook, Sidney
Kang, Frances
Kang, Younghill
Kunitz, Joshua
Lerner, Max
Levey, Jeffrey
Mellquist, Jerome

Algren, Nelson
Asch, Nathan
Bacon, Glenn
Ballou, Jenny
Boardman, Helen
Botkin, Benjamin Albert
Byer, Alexander
Campbell, Gladys Irene
Cheever, John.
Cohen, Morris Raphael
Daly, James Joseph
Dickinson,
Thomas Herbert
Driggs, Elsie
Ehrlich, Leonard
Elwell, Herbert
Farrell, James T.
Fletcher, John Gould
Gatch, Lee
Gregory, Horace
Gruening, Martha.
Haggin, B. H. *Bernard H.*
Hechenbleikner, Christl
Joffe, Eugene
Kishore, Penina

Arent, Arthur
Asch, Nathan
Ascoli, Max
Bermann, Richard A.
Blitzstein, Marc
Broch, Hermann
Buckman, Gertrude
Castellon, Federico
Cheever, John.
Clark, Eleanor
Cohn, Arthur
Collins, Loyd
De Pereda, Prudencio
Diamond, David
Dougherty, Marshall
Ehrlich, Leonard
Fuchs, Daniel
Fuchs, Susan
Gummo, Blanchard
Gumpert, Martin
Hechenbleikner, Christl
Hechenbleikner, Louis
Herbst, Josephine
Hoult, Norah
Jacob, Dr. E.E.

Asch, Nathan
Barker, Shirley
Calmer, Alan
Cantor, Eli
Castellon, Federico
Cheever, John
Cohen, Miriam
Cohn, Arthur
Diamond, David
Ehrlich, Leonard
Engstrand, Sophia Belzer
Engstrand, Stuart David
Evans, Robert Allison
Graf, Oskar Maria
Greenwood, Marion
Guck, Edna
Gummo, Blanchard
Hicks, Granville
Joffe, Eugene
Jones, Paul Haller
Lewis, Flannery
Lockwood, Normand
Long, Arthur
Martin, Peter
Mitchell, Bruce

Allison, Mrs. John
Bates, Ralph
Brinnin, John Malcolm
Burke, Selma
Capote, Truman
Cartier-Bresson, Henri
Cayton, Horace R.
Citron, Minna
Cole, George
Cowley, Malcolm
de Marco, Jean
de Marco, Mrs. Jean
Domino, Ruth
Doughty, Howard, Jr.
Ehrlich, Leonard
Fowlie, Wallace
Garrigue, Jean
Glemser, Bernard
Hackett, Francis
Ishigaki, Eitaro
Ishigraki, Haru
Kay, Ulysses
Lerman, Leo
Mangione, Jerre
Milano, Paolo

Barrett,
Barrett,
Blazek,
Bontem
Brinnin,
Cartier-
Cartier-
Cayton,
Cowley,
Dodson
Domino
Doremu
Dought
Fowlie,
Friend,
Goff, E
Haleff,
Hart, A
Houser,
Kay, Ul
Krim, S
Lasker,
Lemelin
Lenson,
Lowell,

Fig. 2. Yaddo Archives Project: Visit overlaps and relationships in table and timeline format

Yaddocast

In addition to the Yaddo Archive Project, the exhibition also led to the creation of another digital humanities project entitled Yaddocast. Yaddocast is an enhanced podcast series on Yaddo artists. Yaddocast was created to tell the stories behind Yaddo's artist guests and their creative processes. Each episode of Yaddocast was produced as a scholarly investigation into creative activity associated with Yaddo, and the project's utilization of a popular form of web-based media will be addressed as part of its pedagogical aim toward fostering cultural literacy.

Fostering Cultural Literacies through Yaddo-related Digital Scholarship

Multimodal humanities projects can be designed to take advantage of the complex database logics and immense bodies of knowledge related to the history of Yaddo. The cultural impact of Yaddo far exceeds the 5,500 artists who have been guests in Saratoga Springs. Artists associated with Yaddo produce cultural works that influence other cultural producers, and members of the Yaddo interact with one another in novel and surprising ways. YAP, for example, utilizes the latest techniques to make those connections and relationships manifest. Rather than publishing one scholar's take on artistic cross-pollination at Yaddo, YAP is a dynamic hypermedia system that can be reconfigured based on different searches and metadata criteria. Moreover, YAP is a read-write database, where knowledgeable users or Yaddo artists can add further relevant information.

Second, Yaddo requires a rich media approach to represent its range of cultural artefacts. Since Yaddo is not a doctrinaire artist colony, it supports all different kinds of artworks, aesthetic techniques and artistic styles. Yaddo's creative output and the richness of its archives are uniquely positioned to show a broad view of American culture since 1926. But that broad purview also requires innovative approaches to representing a variety of media types and forms in digital scholarship. This paper will address some of the techniques used in these two projects.

Third, YAP and Yaddocast engage in a hermeneutics of creativity activity. The relationship between information, form and aesthetics all matter in these multimodal humanities projects. Rather than just being bits of cultural data, the process behind the creative output and the relationships among Yaddo artists are important vectors of meaning, and part of the process of cultural literacy is not just an examination of the end products (books, films, dances, etc) but a more thorough understanding of the usually hidden aspects of the creative process. The artistic legacy of Yaddo's guests spans almost every art field, and allows for a richness of content that is unmatched by even large museums. But unlike normal museum collections, Yaddo's archives are tied to a specific space of creative endeavor. That linkage is useful from a multimodal humanities development perspective since it allows for a new set of recombinant cultural possibilities relating to how Yaddo operates as a social network of artists.

Taken as a whole, the digital humanities projects related to Yaddo begin to suggest that, within the digital humanities, there are opportunities for fostering cultural literacies through collaborating with larger publics around web-based forms of scholarship that utilize architectures of participation.

References

Bordieu, P. (2002). "The Forms of Capital," in *The Sociology of Economic Life*, Mark S. Granovetter and Richard Swedberg, eds. Cambridge, MA: Westview Press, pp. 96-111.

DeCarlo, D. (2004). *Extreme Project Management*. New York: Wiley-Jossey.

Lin, N. (2001). "A Network Theory of Social Capital," in Social Capital: Theory and Research, Nan Lin, Karen S. Cook, and Ronald S. Burt, eds. New Brunswick, NJ: Transaction Publishers.

McGee, Micki, ed. (2008). *Yaddo: Making American Culture*. Columbia University Press.

McPherson, T. quoted in Spiro, L. (2008). "Doing Digital Scholarship." http://digitalscholarship.wordpress.com/2008/08/11/doing-digital-scholarship-presentation-at-digital-humanities-2008/ (Accessed 14 November 2008)

O'Reilly, T. (2004). "The Architecture of Participation." http://www.oreillynet.com/pub/a/oreilly/tim/articles/architecture_of_participation.html (Accessed 14 November 2008)

Unsworth, J. (2007). "Scholarly Primitives: what methods do humanities researchers have in common, and how mght our tools reflect this?" http://jefferson.village.virginia.edu/~jmu2m/Kings.5-00/primitives.html (Accessed 14 November 2008)

Ask Not What Your Text Can do For You. Ask What You Can do For Your Text (a Dictionary's perspective)

Carlos Monroy
Texas A&M University
cmonroy@csdl.tamu.edu

Richard Furuta
Texas A&M University
furuta@csdl.tamu.edu

Filipe Castro
Texas A&M University
fvcastro@tamu.edu

1. Introduction

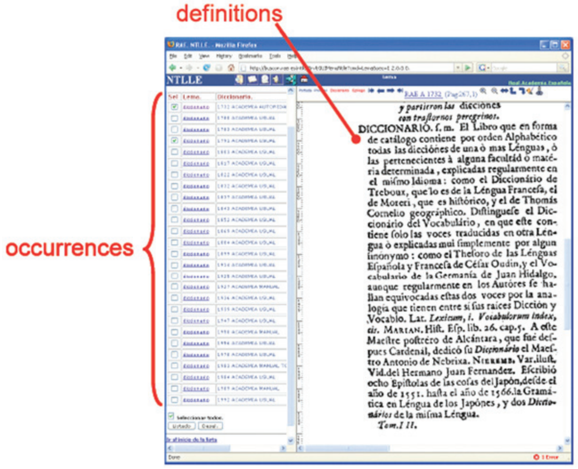
This paper's title, a modified version of President Kennedy's well known quote, in a metaphorical sense suggests shifting the role texts play in a collection. This shift is based on the growing number of available tools that can be used to enhance textual analysis. We are investigating the effect of tools as generators of what can be done with the texts. We are not advocating a conceptual change in the role and nature of the texts from the literary or textual studies perspective. Rather we suggest a pragmatic role change. In doing so, we believe that new hypothesis about the content of the texts can be posited, or at least their use can be augmented.

Our motivation is based on our experience in the creation and use of a multilingual glossary of nautical terms for the Nautical Archaeology Digital Library (Monroy et al, 2006). In this context, we have seen two major benefits from the glossary that affect both the scholarly practices and the collection. First, it has enabled collaboration among scholars and researchers geographically scattered. Second, how it has broadened the possibilities in the use and understanding of the textual materials—shipbuilding treatises in our case.

2. Dictionaries

Dictionaries have been used extensively in numerous digital humanities initiatives. The Perseus Project (Crane 2002) provides a good example of incorporating dictionaries in a classics collection. The idea behind a dictionary is very simple: an alphabetical list of words with definitions. Yet it has great potential and usefulness when used simultaneously with the contents of a digital collection. Dictionaries also come in various flavors—bilingual or multilingual, thesaurus, specialized, illustrated, and encyclopedic—to name a few. With the use of information technology and the Internet, it has been possible to expand not only their use, but also the way they are created and edited.

Searching for a term in the on-line dictionary of the Real Academia de la Lengua Española (RAE, 2008), for example, presents users with occurrences of that term in all the digitized dictionaries where it has ever been edited; see Fig. 1. Because all editions of the printed version of the dictionary have been digitized, it is possible to visualize the evolution of the definitions of a given term. Although this electronic version of the dictionary in itself is a great resource, one can imagine what could be accomplished if used in combination with a corpus of texts.

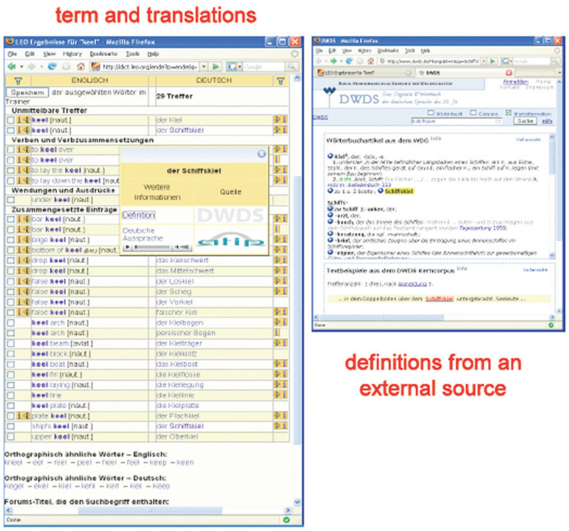


y partirron las dicciònes
con traſtornos peregrinos.
DICCIONARIO. ſ. m. El Libro que en forma de catálogo contiene por orden Alphabético todas las dicciónes de una ò mas Lénguas, ò las pertenecientes à alguna facultád ò matéria determinada, explicadas regularmente en el miſmo Idioma: como el Diccionário de Treboux, que lo es de la Léngua Franceſa, el de Moreri, que es hiſtórico, y el de Thomás Cornelio geográphico. Diſtinguese el Diccionário del Vocabulário, en que eſte contiene ſolo las voces traducidas en otra Léngua ò explicadas mui ſimplemente por algun ſinónymo: como el Theſoro de las Lénguas Eſpañola y Franceſa de Céſar Oudin, y el Vocabulario de la Germanía de Juan Hidalgo, aunque regularmente en los Autóres ſe hallan equivocadas eſtas dos voces por la analogía que tienen entre sí ſus raices Dicción y Vocablo. Lat. *Lexicum, i. Vocabulorum index, cir.* MARIAN. Hiſt. Eſp. lib. 26. cap. 5. A eſte Maeſtre poſtréro de Alcántara, que fué deſpues Cardenál, dedicó ſu *Diccionário* el Maeſtro Antonio de Nebrixa. NIEREMB. Var. iluſt. Vid. del Hermano Juan Fernandez. Eſcribió ocho Epiſtolas de las coſas del Japón, deſde el año de 1551. haſta el año de 1566. la Gramática en Léngua de los Japónes, y dos *Diccionários* de la miſma Léngua.
Tom. I II.

Fig. 1 A screen shot of the on-line RAE dictionary depicting occurrences of a term in the collection of digitized dictionaries, on the right is the image of a given occurrence

Arachne (Foertsch 2006) is an electronic repository (database) of the German Institute for Archaeology. Because archaeological objects are scattered across the world, Arachne provides multilingual access and thesaurus. The Getty Thesaurus of Geographical Names (Baca 2004) is another good example of an external tool that can be incorporated into existing textual materials, enhancing searching and browsing.

The LEO on-line dictionary (LEO, 2008) was originally launched as a German-English dictionary in 1995. At present it includes German translations into French, Spanish, Chinese, and Italian. Since its beginning, two of the most remarkable accomplishments have been the integration of a larger and linguistic diverse editorial team. And the creation of new environments for searching, using, and learning. For instance, the current version allows users to join groups and work together to learn the language; it also enables teachers to organize lessons, see Fig. 2.



ENGLISCH	DEUTSCH
Speichern der ausgewählten Wörter im Trainer	29 Treffer
Unmittelbare Treffer	
keel [naut.]	der Kiel
keel [naut.]	der Schiffskiel
Verben und Verbzusammensetzungen	
to keel over	
to keel over	
to lay the keel [naut.]	
to lay down the keel [naut.]	
Wendungen und Ausdrücke	
under keel [naut.]	
Zusammengesetzte Einträge	
bar keel [naut.]	
bar keel [naut.]	
bilge keel [naut.]	
bottom of keel [naut.]	
drop keel [naut.]	das Kielschwert
drop keel [naut.]	das Mittelschwert
false keel [naut.]	der Loskiel
false keel [naut.]	der Scheg
false keel [naut.]	der Vorkiel
false keel [naut.]	falscher Kiel
keel arch [naut.]	der Kielbogen
keel arch [naut.]	persischer Bogen
keel beam [aviat.]	der Kielträger
keel block [naut.]	der Kielklotz
keel boat [naut.]	das Kielboot
keel fin [naut.]	die Kielflosse
keel laying [naut.]	die Kiellegung
keel line	die Kiellinie
keel plate [naut.]	die Kielplatte
plate keel [naut.]	der Flachkiel
ship's keel [naut.]	der Schiffskiel
upper keel [naut.]	der Oberkiel

Orthographisch ähnliche Wörter – Englisch:
kneel – eel – feel – peel – heel – reel – keep – keen

Orthographisch ähnliche Wörter – Deutsch:
Kegel – ekel – Kiel – kehl – Ken – Kel – Keep

Forums-Titel, die den Suchbegriff enthalten:

Wörterbuchartikel aus dem WDG

Schiffs-

Textbeispiele aus dem DWDS Kerncorpus

... in dem Doppelboden über dem Schiffskiel untergebracht. Seeleute ...

Fig. 2 A screen shot of a bilingual dictionary—LEO—depicting translations and definitions

3. Motivation

Although the use of tools for textual analysis is not a new concept in digital humanities, their emergence is reshaping not only the use of textual materials, but also the mere notion of the texts themselves. Impacting in the end how scholarly practices are conducted.

Traditionally, textual scholars can be seen as "consumers" of the texts. They analyze, compare, and study their contents, how they relate to each other, and their historical and cultural contexts to name a few. Rockwell (2003) commenting on a well known discussion about two approaches to texts: as a hierarchical objects advanced by Renear, and as a performance proposed by McGann, states:

> If we are to take McGann's public performance of a reading as an analogue for what we wish to achieve with these tools, we have to think not only about how we represent the text but also about the performance of analysis and the tools that are used to perform this analysis with a computer.

In this paper, following Rockwell's statement, we describe a change in the role of textual scholars, from *consumers* (users of the texts) into *producers* (augmenting the texts) with the use of *tools*. Our observations are based on the function external tools can play in augmenting the use of the texts; how they can be used; and what can be learned. Therefore, our goal is neither to propose a right approach, nor to compare approaches.

We take this approach for two reasons. First, at a recent textual studies conference (CASTA, 2008) a participant asked one of the presenters regarding the numerous tools available to text scholars: "*But don't you think that quite often the problem with 'tools,' is precisely that there are too many of them; and we don't know what to do?*" This is an interesting question because—although coming from a literary scholar with strong background and expertise in using technology for textual studies—it shows the marked prominence of the role texts play in digital humanities, or at least in how humanists perceive their role.

The second reason is based in our experience working in the creation of a multilingual glossary of nautical terms. The glossary has allowed the incorporation of a new layer to the original transcriptions. For example it would be possible to search for a given term in one language and retrieve occurrences in the transcriptions in multiple languages. Also categories associated to the terms can be used for retrieving occurrences in various contexts.

Used in the context of the Willa Cather Archive (The Willa Cather Archive, 2008), Evince—a non-invasive text analysis tool that mediates the integration of analytical data with the text—shows how a tool can enhance the textual materials. What is interesting in this case, is the fact that the tool is being used to augment the study of the texts, hence improving what can be learned from them. Discussing the use of Evince, (Jewell et al., 2008) state:

> We posit that integration of analytical data with the reading text will create new possibilities for interpretation informed by textual data, as it will eliminate the need to enter a specialized environment.

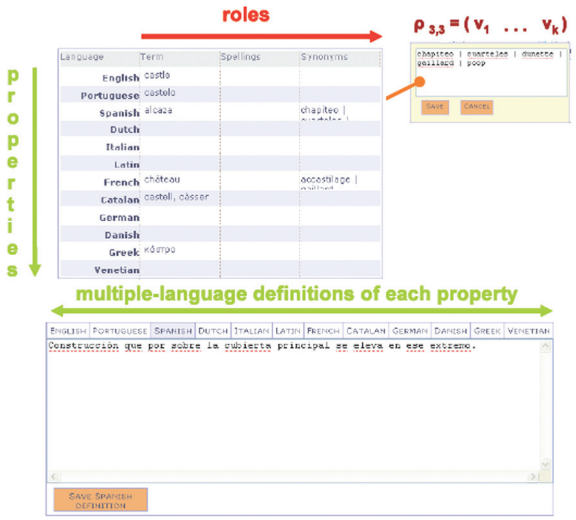
4. Our Collection of Shipbuilding Treatises

Shipbuilding treatises are ancient technical texts, both printed or manuscript, that describe the conception and construction of ships, establish the required types and properties of the wood and building materials utilized, and sometimes describe the steps to be followed in their construction. Given their characteristics, these texts can be properly considered as ancient technical manuals. Our collection was started with three Portuguese treatises obtained with permission from the Portuguese Academia de Marinha and National Library. At present our collection has grown to eleven copies; and includes materials in Portuguese, French, Italian, and Dutch, spanning a period from the late 16th to the early 18th centuries. Additionally, an English book is already digitized and ready to be added.

In terms of naval and seafaring dissemination, shipbuilding treatises are priceless sources for scholars working in ship reconstruction and studying the evolution of shipbuilding techniques. Moreover, the development of underwater archaeology in the last 50 years propitiated the growth of the archaeological data corpus, which can now be tested against the textual evidence pertaining to the conception and construction of these complex machines. Nautical Archaeology students, on the other hand, study ship treatises as part of their curriculum. Finally, for the general public they are a great source of historical and cultural contexts in which seafaring flourished.

5. The Tool—A Multilingual Glossary of Nautical Terms

The need for the creation of our glossary goes back to an English illustrated glossary included as appendix in an underwater archaeology book (Steffy, 1994). Tied to one language and a printed medium, the glossary's limitations were evident. But the most pressing reason was the various languages in which the texts in our collection were written. Further, the glossary is essential because nautical archaeology is a highly specialized domain where technical terms need to be explained in order to understand their meaning and context.


roles
$\rho_{3,3} = (v_1 \ \ldots \ v_k)$
properties
multiple-language definitions of each property
SAVE
CANCEL
SAVE SPANISH DEFINITION

Language	Term	Spellings	Synonyms
English	castle		
Portuguese	castelo		
Spanish	alcaza		chapiteo |
Dutch			
Italian			
Latin			
French	château		accastilage |
Catalan	castell, càsser		
German			
Danish			
Greek	κάστρο		
Venetian			

chapiteo | cuarteles | dunette | gaillard | poop

ENGLISH | PORTUGUESE | SPANISH | DUTCH | ITALIAN | LATIN | FRENCH | CATALAN | GERMAN | DANISH | GREEK | VENETIAN

Construcción que por sobre la cubierta principal se eleva en ese extremo.

Fig. 3 A partial display of the NADL glossary depicting our model to represent a multilingual dictionary of nautical terms

Our model uses *term* as the atomic element. Each term in turn has an associated matrix where columns correspond to *roles* and rows to *properties*. Because we are working on a multilingual glossary, we decided to use *properties* to map languages, while *roles* map synonyms and spellings respectively, see Fig 3. Each cell at the intersection of role and language can contain zero, one, or more values separated by the symbol |. This implies that each cell can be represented as a vector of values.

Our approach allows scalability and flexibility. For example, we had to add a new language—Venetian—since it was not originally considered, and was requested by one of the scholars. Adding the new language was a straightforward process. Similarly, adding new *roles* entails the addition of more columns. In both cases both the architecture and the interface scale easily. From the implementation standpoint, we use a relational database for storing terms, synonyms, spellings, and definitions in multiple languages.

Using Lucene—an open source full-text retrieval software (The Apache Lucene Project, 2008)—we are parsing texts and automatically creating links to the glossary. Fig.4 depicts a screen shot of the treatises interface showing the image on the left, and the transcription on the right, with linked terms underlined in blue. Although this process might seem a simple one, implementing it turned out more complicated than expected. The two main reasons were multiple-word entries and the limitations on Lucene's stemmer to handle 17th-century Portuguese.

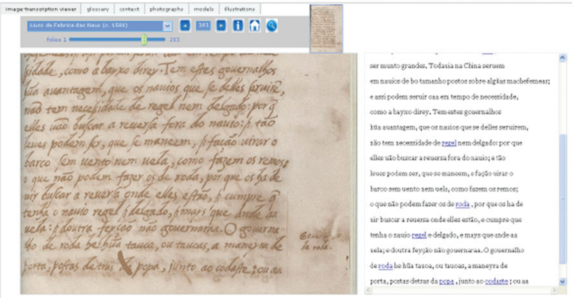


image-transcription viewer
glossary
context
photographs
models
illustrations
Livro da Fabrica das Naus (c. 1580)
161
folios 1
213

ser munto grandes. Todauia na China seruem
em nauios de bo tamanho postos sobre algũas machefemeas;
e assi podem seruir caa em tempo de necessidade,
como a bayxo direy. Tem estes gouernalhos
hũa auantagem, que os nauios que se delles seruirem,
não tem necessidade de rezel nem delgado: por que
elles não buscar a reuersa fora do nauio; e tão
leuos podem ser, que os manoem, e fação uirar o
barco sem uento nem uela, como fazem os remos;
o que não podem fazer os de roda , por que os ha de
uir buscar a reuersa onde elles estão, e cumpre que
tenha o nauio rezel e delgado, e mays que ande aa
uela; e doutra feyção não gouernaraa. O gouernalho
de roda he hũa tauoa, ou tauoas, a maneyra de
porta, postas detras da popa , junto ao codaste ; ou as

Fig. 4 The treatises interface depicting linked terms in the transcriptions

6. Conclusion

Our Web-based interface has enabled scholars to work remotely in editing the glossary, expanding its contents and attracting other scholars. This collaboration goes beyond merely the editing of materials remotely. It has allowed us to obtain materials from other libraries and also to engage the special collections library at Texas A&M in the acquisition of original materials.

Profiting from the rich illustrations nautical treatises provide and the numerous ship models in our collection, we want, in the near future, to create multilingual illustrated dictionaries, linking them to the texts. As stated earlier, our goal is not to redefine the role of texts in the humanities. But as our experience with the introduction of the multilingual glossary in NADL indicates, tools are shifting the way texts are perceived.

7. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0534314.

8. References

Baca, M. (2004). *Fear of Authority? Authority Control and Thesaurus Building for Art and Material. Cataloguing & Classification Quaretly* Vol. 38, No. 3/4, pp. 143-151.

Crane, G. (2002). *Cultural Heritage Digital Libraries: Needs and Components*. In ECDL 2002, LNCS 2458, pp. 626-637, 2001. Springer-Verlag, Berlin, 2002.

Foertsch, R. (2006). *ARACHNE - Datenbank und kulturelle Archive des Forschungsarchivs fuer Antike Plastik Koeln und des Deutschen Archaeologischen Instituts*. http://arachne.uni-koeln.de/ (accessed 12 October 2008).

Jewell, A., Zilig, B., and Ramsay, S. (2008). *Can Text Analysis Be Part of the Reading Field?: The Vision of Evince*. CaSTA 2008, New Directions in Text Analysis. http://ocs.usask.ca/ocs/index.php/casta/casta08/paper/view/25 (accessed 2 November 2008).

Monroy, C., Parks, N., Furuta, R., and Castro, F. (2006). *The Nautical Archaeology Digital Library, 10th European Conference on Research and Advanced Technology for Digital Libraries ECDL,* Alicante, Spain, September 2006. In Gonzalo et al. (Eds.) - LNCS 4172 :544-547, Berlin and Heidelberg: Springer-Verlag, 2006.

Monroy, C., Furuta, R., and Castro F. (2007). *A Multilingual Approach to Technical Manuscripts: 16th and 17th-century Portuguese Shipbuilding Treatises. ACM-IEEE Joint Conference on Digital Libraries*, Vancouver, British Columbia, Canada, June 2007.

Rockwell, G. (2003), *What is Text Analysis, Really?* Literary and Linguistic Computing 18(2):209-219, 2003.

Steffy, D. (1994). *Wooden Ship Building and the Interpretation of Shipwrecks*. Texas A&M University Press, College Station, Texas (1994).

Diccionario de la Real Academia de la Lengua Española, http://www.rae.es/rae.html (accessed 28 October 2008).

CASTA 2008 New Direction in Text Analysis, http://ocs.usask.ca/ocs/index.php/casta/casta08/index/ (accessed 17 October 2008).

The Apache Lucene Project, http://lucene.apaches.org/ (accessed 10 October 2008).

The LEO dictionary, http://dict.leo.og/ende?lang=en (accessed 30 October 2008).

The Willa Cather Archive, University of Nebraska Lincoln http://cather.unl.edu (accessed 1 November 2008)

Visualizing Archival Collections with ArchivesZ

Jeanne Kramer-Smyth
University of Maryland, College Park
jeanne@spellboundblog.com

Jennifer Golbeck
University of Maryland, College Park
jgolbeck@umd.edu

Archival records and manuscripts are usually arranged to retain their original order when transferred into the care of archivists or manuscript curators. A side effect of grouping records by record creator and retaining the creator's original organization is that materials are described at the group level—not at the item level. The ramifications for archive searchability are dramatic: Imagine a library where, instead of being grouped together by subject, books were shelved alphabetically by author—and interspersed with each author's notes, drafts, expense records, and personal memorabilia. This basic difference between libraries and archives is key to understanding why subject-based access to archival resources is both challenging to achieve and very useful when available.

With this in mind, we have developed ArchivesZ, an information visualization tool for archival collections. It enables users to visualize and explore aggregated information relating to the total linear feet, inclusive years and subject terms for archival collections extracted from EAD encoded finding aids.

Chris Anderson of *Wired* described the power of the "long tail" in his *Wired* article of the same name. He discussed that the future belonged not to the bestsellers, but rather to "the millions of niche markets at the shallow end of the bit-stream."[1] There has been much discussion of the long tail with regard to library resources[2], but it is interesting note that archival materials are virtually all long tail. The nature of archival collections is such that many of those with the greatest desire to access the materials have very narrow and specific interests. It is quite rare that the documents in a single archival collection will be popular, in the sense of a bestselling book. Frequently it is a challenge for humanities scholars wishing to use archival materials to figure out how to approach the search process. Use of a visualization tool designed to support the examination of aggregated information about archival collections could support a more serendipitous process of exploration of materials and the discovery of new avenues of research.

Even the most experienced historian or humanities scholar has struggled with the challenge of locating relevant primary sources. Archival record groups and manuscript collections present unique challenges to researchers. For example, a standard search result list shows only the title and short description for each record group or collection. This list fails to convey the quantity of materials or diversity of subjects covered by the combination of collections returned by the search. A visualization tool that supports examination of cross-collection and cross-institution aggregated data about archival collections could:

- Encourage the browsing and exploration of locally available cultural heritage resources,
- Improve understanding of existing collections,
- Permit easy identification of locations with a rich combination of collections applicable to a particular research project, and
- Increase interest in both the humanities and primary materials.

Encoded Archival Description (EAD) is the international de facto standard for encoding archival finding aids in an XML format. Finding aids include information about who created the records, when they were created, why they were created, what topics the records relate to, and the size of the collection. The archival community has spent much of the past decade encoding existing finding aids using the EAD standard. Up to this point the major selling point of EAD has been as a tool for simplifying the process of publishing finding aids online. While work has been done to create tools to facilitate the encoding of finding aids, the next step is to take advantage of the structured data now available in EAD encoded finding aids. This machine readable data can support the creation of innovative software programs intended to extract, organize, facilitate discovery of and aggregate information about archival resources.

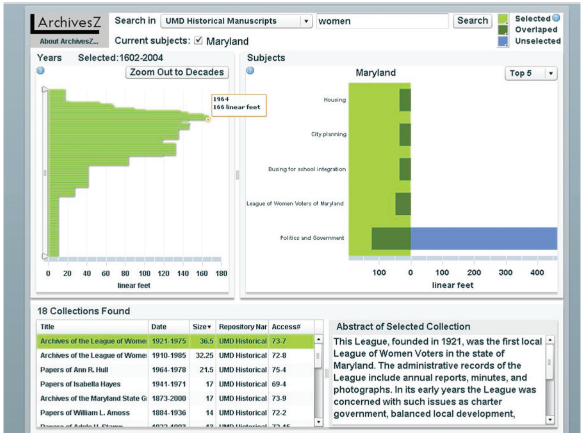
Tools for visualizing archival collections support the needs of three distinct user groups.

- Archivists and manuscript curators can use such a tool to improve their understanding and validate the metadata of the collections at various institutions including their own.
- Literary researchers, historians and humanities scholars can use this type of tool to permit easy

identification of institutions with archival collections fitting the criteria of their research.

- Finally, this type of tool can enable exploration of locally held cultural heritage materials by students and promote use of primary sources. In contrast to researchers who frequently have very specific interests before they examine the collections held by an institution, students in the university setting who are interested in humanities topics are likely not aware of the primary sources available. A tool of this type might encourage the browsing and open ended exploration of locally available cultural heritage resources, and increase interest in both the humanities and primary materials.

Built in spring of 2007, the first version of ArchivesZ is a prototype for just such a tool. Designed to support search, exploration and visualization of archival record groups and manuscript collections, ArchivesZ addresses a major challenge facing humanities scholars - the need to understand the scope and quantity of available archival records and manuscripts.



18 Collections Found

Title	Date	Size▾	Repository Nar	Access#
Archives of the League of Wome	1921-1975	36.5	UMD Historical	73-7
Archives of the League of Wome	1910-1985	32.25	UMD Historical	72-8
Papers of Ann R. Hull	1964-1978	21.5	UMD Historical	75-4
Papers of Isabella Hayes	1941-1971	17	UMD Historical	69-4
Archives of the Maryland State G	1873-2000	17	UMD Historical	73-9
Papers of William L. Amoss	1884-1936	14	UMD Historical	72-2

Abstract of Selected Collection

This League, founded in 1921, was the first local League of Women Voters in the state of Maryland. The administrative records of the League include annual reports, minutes, and photographs. In its early years the League was concerned with such issues as charter government, balanced local development,

Fig. 1 Screenshot of ArchivesZ Prototype (video demonstration online at archivesz.org)

To support organic exploration of subject terms associated with collections, ArchivesZ leverages a unique dual sided histogram (see right half of Figure 1). The ArchivesZ prototype combines this dual sided histogram with a more traditional histogram displaying year data to permit tightly coupled, multi-dimensional browsing of subject and time period metadata. By representing the distribution of subjects and time periods using the metric of total aggregate linear feet, ArchivesZ permits users to get a better sense of total available research materials than they would by viewing a standard search result list. The subject term visualization interface may also support a deeper understanding of the relationships among subject terms through the lens of the currently selected set of collections.

Further development of ArchivesZ has been supported by a National Endowment for the Humanities Digital Humanities Startup Grant. In this Poster / Demo, we will present the newest version of ArchivesZ in use over a large set of finding aids provided by a wide range of partner archives. Our demo will show the newest version of he tool and we will discuss how this lays the foundation for the future creation of a public tool for visualizing archival collections.

Notes

[1]C. Anderson. The long tail. *Wired*, 12(10), 2004.

[2] L. Dempsey. Libraries and the long tail: Some thoughts about libraries in a network age. *D-Lib Magazine*, 12(4), April 2006.

Active Animation: An Approach to Interactive and Generative Animation for User-Interface Design and Expression

Kenny K. N. Chow
Georgia Institute of Technology
knchow@gatech.edu

D. Fox Harrell, Ph.D.
Georgia Institute of Technology
fox.harrell@lcc.gatech.edu

Abstract

The traditional view of animation is a medium-specific perspective: animation is a sequence of images on film. In contrast, we employ a wider, interdisciplinary theoretical lens, based on a phenomenological perspective of animation. We describe animation as the experience of artifacts imbued with apparent "animacy," or "liveliness," and identify a range of media artifacts where an account of animacy is key to understanding and designing their functionality. These artifacts range from computer interface mechanisms such as bouncing and stretching icons to interactive cartoons that may be used for informational, entertainment, or socio-critical purposes. We introduce the term "active animation" to describe this range of artifacts. Insights from textual analyses in the humanities-based field of animation studies can enable analysis and design of active animation, and likewise animation studies can be informed by insights regarding agency in artificial intelligence research, theories of embodied cognition and conceptual blending from cognitive science, and psychological approaches to movement and perception. To exemplify the technical design potential of our approach, we present a cognitive semantics-based interactive and generative multimedia system that we have implemented called the Generative Visual Renku system as a case study active animation. The upshot is that our interdisciplinary animacy-oriented perspective highlights how gesture and movement allow interactive and generative digital artifacts to convey non-verbal meaning to users.

1. Introduction

Animacy lies at the heart of many media artifacts imbued with an illusion of life. Puppets and avatars, examples of traditional and digital user manipulated characters, become lively under the control and enactment of performers. Animatronic robots utilize mechanical means to produce the appearance of gesturing and perceiving viewers. Cartoons, manifested through sequences of pictures, can walk like real human figures. Although these artifacts differ from each other in terms of material form, control mechanism, and technology, all of them are animated in a literal sense. The animation of these artifacts hinges upon lively motion as the primary phenomenon of illusion of life.

Meanwhile, the term animation is often narrowly seen as referring to a particular medium, namely a type of film. Indeed, the celebrated filmmaker Norman McLaren describes animation as the "art of movements that are drawn." (Wells, 1998) Although his quote seemingly privileges motion over medium, the material condition of imagery as drawings is still presumed. In contrast, we call attention to views that deemphasize medium and emphasize liveliness. The animation theorist Alan Cholodenko attempts to generalize the notion of animation as sorts of technology geared toward "endowing with life" and "endowing with motion." (Cholodenko, 2007) In parallel, many digitally mediated environments such as computer interfaces, websites, and handheld devices have become lively, reactive, semi-autonomous, and graphical. They often construct meaning through perceived movement and embodied interaction. We call digital images engaged in such meaning-making processes "active animation." Given the ubiquity of such multimedia computing phenomena that are often overlooked as animation, there is need for theory to comprehend how such artifacts convey non-linguistic meaning via animacy and to formulate theoretically-grounded approaches for designing lively multimedia artifacts. This paper articulates this need and presents a new approach to addressing it, including a new form of active animation that we have developed.

2. Theoretical Framework

Our approach to the analysis and design of active animation arises from an intersection of multiple disciplines. Animation and image studies provide us a critical vocabulary for identifying the phenomenon of liveliness as definitional for our area of inquiry. (Arnheim, 1974; Metz, 1974; Mitchell, 1986) For thinkers such as Ludwig Wittgenstein and W.J.T. Mitchell, the term "image" is not limited to material images (e.g. screen images) or optical images, but also means perceptual images (through motor-sensory functions, including the kinaesthetic) or even mental images. It follows that the idea of animation should also extend to considering moving images on the basis of sensory perception and embodied cognition. Cognitive semantics research provides accounts of embodied meaning construction and generation of imaginative meaning through metaphorical projection and conceptual blending. (Fauconnier & Turner, 2002; Lakoff & Johnson, 2003; Turner, 1996) Psychological and phenomenological approaches to human-computer interaction are also relevant departure points for investigating the role of interactivity in the perception of liveliness. (Norman, 1988; Shneiderman, 2003)

3. Active Animation: Examples and Analyses

Toward analyzing and designing instances of active animation we introduce two levels of signification: the reactive and the metaphorical. At the reactive level, users make meaning out of liveliness of artifacts through a motor-sensory loop feedback between users and systems – i.e. users perform actions via an interface and perceive their animated effects in the system. Examples include user-interface mechanisms such as the many shrinking, stretching, and bouncing icons in the Macintosh OS X environment and interactive animated comics such as found at www.hoogerbrugge.com.

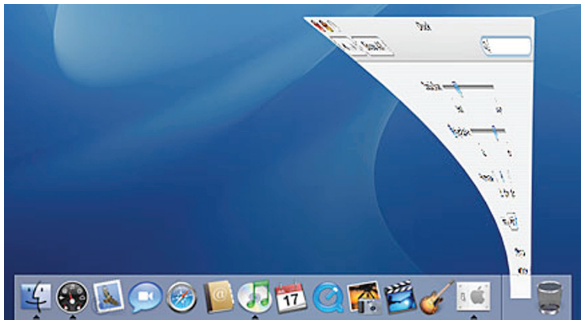
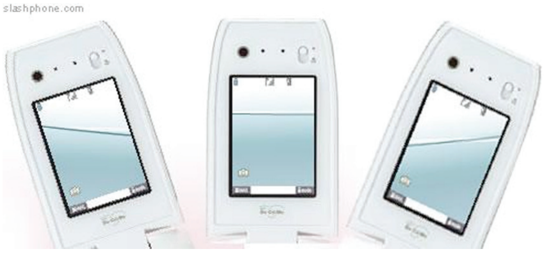


Fig. 1 The "genie" effect in Macintosh OS where windows dynamically stretch and shrink

Such works imbue media elements such as windows and icons with a sense of liveliness formerly unknown in user interfaces. Motion is used to focalize user attention, add spectacle to basic operation, and to allow embodied user action such as clicking to play a role in realizing the meaning of animated content. (Arnheim, 1974; Lakoff & Johnson, 1999) Basic image schemata (skeletal patterns of motor-sensory perception) play crucial roles in user understanding of such works. (Lakoff & Johnson, 2003) For example, the "dock" area of the Macintosh graphical user interface becomes a container for windows, paralleling the container image schema articulated in (Lakoff & Johnson, 2003).

At the metaphorical level, users construct imaginative motion metaphors through the interaction between embodied gestures and multimedia feedback. The idea can be demonstrated by the water-level interface designed for the mobile phone N702iS and the electronic advertising viral campaign www.comeclean.com.



slashphone.com

Fig. 3 A mobile phone interface where battery level is indicated via the illusion of a water-filled container

The water-level interface in **Fig. 3** comprises a conceptual metaphor in which a container filled with water is integrated with a standard interface element depicting battery-level. This metaphor exploits the liveliness of animated water to present functional information in a lively and playful manner. The website Comeclean.com invokes standard interface mechanisms such as data-entry and mouse-clicking to arrive at a metaphorical projection in which users can wash away the wrongdoings. The site is, in fact, an advertisement for cleaning supplies, yet the metaphorical mapping from washing one's hands using particular cleaning supplies to washing away confessions of sin is enabled by the active animation interface.

4. A New Form of Active Animation: Generative Visual Renku

As an example of multimedia system design based on our approach to active animation briefly we present an expressive project that we have developed called Gen-

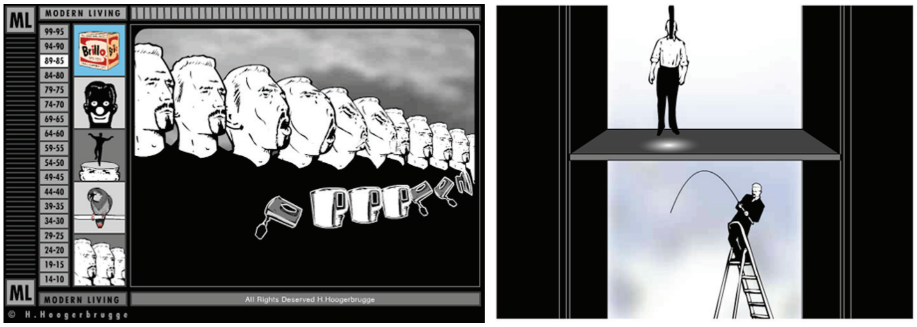


Fig. 2 Two screenshots from Hoogerbrugge.com where interaction drives provocative animated reaction

erative Visual Renku. (Harrell & Chow, 2008) A polymorphic poem is a generative digital artwork that is constructed differently upon each instantiation, but can be meaningfully constrained according to aspects such as theme, metaphor, affect, and discourse structure. Our Generative Visual Renku project presents a work of active animation as a new form of concrete polymorphic poetry inspired by Japanese renku poetry, iconicity of Chinese character forms, and generative models from contemporary art.

In the Generative Visual Renku project interactive iconic illustrations are conjoined by a cognitive science based computer program called GRIOT into a fanciful topography. GRIOT, which is a system for composing generative and interactive narrative and poetic works, is used to semantically constrain generated animated output both visually and conceptually.

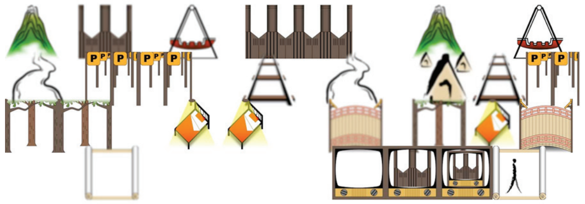


Fig. 5 A Generative Visual Renku screenshot: Users co-create animated maps by clicking visual icons, the system responsively selects subsequent images according to semantic constraints

Conclusions and Implications

Today many multimedia computing systems show spectacular animated images that react to user actions with animated feedback. These artifacts manifest the notion of animation in a new horizon beyond the cinematic. The examples of active animation above illustrate this manifestation in both functional interface design such as the lively windows of Mac OS X and mobile phone water-level interfaces and in expressive works such as found at Hoogerbrugge.com or in our own generative visual renku. These works all evoke senses of liveliness, not only with perceptual movements, but also through generative multimodal feedback loops. They bring life to the computer, it can now feel more intimate to users through perceived emotion and even intelligence.

Active animation "enlivens" the computer by concealing its complexity with a "skin" like the shells of animatronic robots. Careful understanding of how users interpret active animation allows designers to "stage" and "veil" technology in order to create spectacles, suspense, surprise, and intuitive non-verbal meanings for users. This approach also brings concern for humanistic interpretation back to the center of analysis and design of multimedia artifacts. The integration of computational and cognitive research results with approaches from animation studies provides a new orientation for designing technologies that are more in line with our everyday, non-verbal, affective acts of communication and understanding.

References

Arnheim, R. (1974). *Art and visual perception : a psychology of the creative eye*. Berkeley: University of California Press.

Cholodenko, A. (2007). Speculations on the Animatic Automaton. In A. Cholodenko (Ed.), *The illusion of life II : more essays on animation* (pp. 486-528). Sydney, N.S.W.: Power Pub.

Fauconnier, G., & Turner, M. (2002). *The way we think : conceptual blending and the mind's hidden complexities*. New York: Basic Books.

Harrell, D. F., & Chow, K. K. N. (2008). Generative Visual Renku: Linked Poetry Generation with the GRIOT System, *Visionary Landscapes: Electronic Literature Organization 2008 Conference*. Washington State

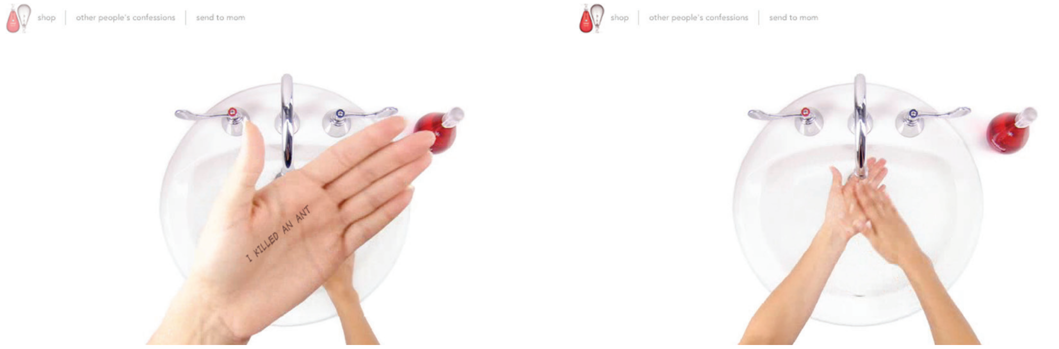


Fig. 4 Screenshots of an active animation web design at comeclean.com where mouse actions cause images of hands to wash away text input by users

University Vancouver, Vancouver, Washington, USA.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh : the embodied mind and its challenge to Western thought*. New York: Basic Books.

Lakoff, G., & Johnson, M. (2003). *Metaphors we live by*. Chicago: University of Chicago Press.

Metz, C. (1974). *Film language : a semiotics of the cinema*. New York: Oxford University Press.

Mitchell, W. J. T. (1986). *Iconology : image, text, ideology*. Chicago: University of Chicago Press.

Norman, D. A. (1988). *The Psychology of Everyday Things*. New York: Basic Books Inc.

Shneiderman, B. (2003 [1983]). Direct Manipulation: A Step beyond Programming Languages. In

N. Wardrip-Fruin & N. Montfort (Eds.), *The new media reader* (pp. 486-498). Cambridge,

Mass. ; London: MIT Press. **Turner, M.** (1996). *The literary mind*. New York: Oxford University Press. **Wells, P.** (1998). *Understanding animation*. London ; New York: Routledge.

I Am a Black Scholar: A Digital Repository of Scholarship from within the Black Diaspora

Leshell Hatley
University of Maryland, College Park
leshell@umd.edu

Introducing The Black Scholars Index™

'I am a Black Scholar' is the slogan for the digital humanities project named The Black Scholars Index™ (BSI) sponsored by Uplift, Inc. Uplift, Inc. is a 501c3 (tax exempt) nonprofit organization, founded by Leshell Hatley, a doctoral student in the College of Information Studies at the University of Maryland. Its mission is to support underrepresented groups and to research and develop technology-based programs, products, and services to encourage and produce lifelong learners, leaders, and resilient communities. With this in mind, the organization and its founder constructed The Black Scholars Index™ (BSI) as an online venue to highlight, record, analyze, and illustrate the scholarly achievements and intellectual movements within the Black Diaspora, descendents of Africans who have settled throughout the world as a result of the Atlantic slave trade. In keeping with current uses of Internet technology, BSI will host features such as: an extensive directory of Black Scholars; *Talented Tenth*, BSI's online Journal for Black Scholars; the BSI Digital Repository, a database of dissertations, research papers, and other work produced by Black scholars; online collaboration tools; a network for mentoring potential students of higher education; pod/webcasts, and a host of Web 2.0 data visualization elements. These tools and features will provide analytical measurements, share insightful perspectives, and enhance the historical account of the knowledge production and of the many societal contributions by Black Scholars.

Investigating The Black Scholars Index™

While many researchers have previously attempted to gather information about the perspectives, experiences, motivations, and challenges of Black students in higher education, few, if any, have done so on large scales, from around the world, nor collected works and experiences of these scholars after they have been working in their fields. In the form of books and articles, some Black students have given personal accounts of their stories at Predominately White Institutions (PWIs) or within fields of study where Black students are underrepresented as attempts to share their experiences and offer advice

(Green 2008; Booker 2007; Hall-Green, 2000; Scott 1995). Over the years, the desire to provide analytical commentary, share knowledge, and provide advice as a result of lived experiences has been prominent themes within the Black intellectual community, especially within the United States. Harold Cruz (1967) provided a historical analysis of what he deemed the challenges of black leadership in his book, *The Crisis of the Negro Intellectual*. In it, he chronicled the major decisions and actions of prominent Black literary, artistic, and political leaders. This book provided an astounding glimpse into Black life similarly to the renowned books of W. E. B. DuBois, *The Souls of Black Folk* (1903) and Carter G. Woodson, *The Mis-education of the Negro* (1933). Nikki Giovanni, a prominent Black literary scholar, dedicated two chapters in her book, *Racism 101* (1994), for advice to African-American students enrolled in college (or soon to be); and in 2003, Anna Green and LeKita Scott published *Journey to the PhD: How to Navigate the Process as African Americans*, a collection of seventeen essay describing the challenges of twenty doctoral student from institution around the United States.

These books are often recommended to members of the Black community and beyond, especially those contemplating higher education, as representations of past accomplishments and challenges from which to gain invaluable lessons. While they are valuable resources, they are not as widely accessible as resources made available on the internet and they do not contain information over extended periods of time. The Black Scholars Index™ provides this and more. Not only will this historical website serve as a repository of research and analytical accounts of knowledge production, it will be an ongoing spotlight of the lived experiences of Black Scholars and offer invaluable advice throughout the life of the project. With it, future scholars of all races will have access to resourceful information, authentic experiences and perspectives, visual representations of research trends, and various other tools and features that will help understand the educational and intellectual activities of Black Scholars.

However, gains from The Black Scholar Index™ existence stretch beyond these benefits. Aside from the project's goals and the above component descriptions, the informative advantages and implications of BSI are tremendous and research has been underway since its inception in February 2008. Two months after launching BSI with only word of mouth marketing, approximately 200 Black Scholars were indexed and initial demographic statistics were measured. Measurements from this snapshot are displayed below:

- Female -66%
- Male -34%.
- In a relationship -22%
- Married -27%
- Single -48%
- Parents – 33%
- Have published papers about their work or similar interests -43%
- Members of fraternities and sororities -21%
- Attended an HBCU for undergraduate degree -49%
- Top BS Degree obtained = BS Biology
- Attended an HBCU for Masters degree -19%
- Top Masters Degree obtained = MS Public Health
- Attended an HBCU for Doctoral degree -3%
- Top PhD Degree obtained = PhD in Computer Science
- Current faculty (tenured and non-tenured) -16%
- Currently doctoral students – 37%
- More than 1/2 work in a field related to their PhD
- 77% expressed interest in serving as BSI mentors

The results of this preliminary snapshot were extremely revealing: the majority of Black scholars indexed studied health and science, a small number were employed in faculty positions, a large majority were interested in mentoring students interested in pursuing advanced degrees, and although approximately half attended Historically Black Colleges and/or Universities (HBCUs) while in undergraduate school, attendance at HBCUs diminishes significantly as higher degrees were sought. Most of the Black scholars indexed to date reside within the United States and the BSI website displays a geospatial map of their employment locations. Although these measurements do not reflect the total number of African-Americans with doctoral degrees, it provides a glimpse of demographic information and illustrates the potential for the amount of information that can be garnered overtime. This information along with more qualitative analysis of the experiences, perspectives, and motivations of those indexed via future website polls, surveys, and other forms of interaction will be beneficial to future educational programs, mentoring organizations, and recruitment efforts by a variety of organizations and institutions.

Conclusion and Future Work

Development of The Black Scholars Index™ is ongoing. Design revisions, expansion of content, and the design and implementation of the many tools including those described above are all underway. One of the intended analysis tools examines the collaboration and co-authorship of research projects and papers. Another tool provides an interface for data entry of well-known African American scholars who have lived over the past century. Further analysis of the all the data collected will help fulfill the mission of the project and provide a concise representation of the scholarly accomplishments of the Black Diaspora.

References

Booker, K. C. (2007). Perception of classroom belongingness among African American college students. *College Student Journal*. 3 (41), 178-186

Du Bois, W. E. B (1903). *The Souls of Black Folk, Republished* (1995) 100th Anniversary Edition, Penguin Group

Giovanni, N. (1994). *Racism 101*. Quill, NY

Green, A. (2008). *A Dream Deferred: The experience of an African American student in a doctoral program in science*. Education Spring 2008, Vol. 128 Issue 3, p339-348

Green, A., Scott, L. (2003). *Journey to the PhD: How to Navigate the Process as African Americans*, Stylus, VA

Hall-Greene, D. (2000). *A Qualitative Study on African American and Caribbean Black Males' Experience in a College of Aeronautical Science*. Unpublished doctoral dissertation, Virginia Tech.

Scott, D. W. (1995). *Conditions related to the Academic Performance of African American Students at Virginia Polytechnic Institute and State University*. Unpublished doctoral dissertation, Virginia Tech.

Woodon, C. G. (1993) *The Mis-education of the Negro*, Republished (1992), UBUS Communications Systems

Digital Editions for Corpus Linguistics: Encoding Abbreviations in TEI XML Mark-up

Alpo Honkapohja
University of Helsinki, Finland.
alpo.honkapohja@helsinki.fi

This poster will present the Digital Editions for Corpus Linguistics (DECL) system for encoding manuscript abbreviations in TEI-conformant XML. First, I will briefly describe the DECL project, its presenting its aims and editorial policies. Secondly, I will go through the problems resulting from the silent expansion of abbreviations, an approach some digital editions derive from traditional editing. And finally, I will describe the possibilities of TEI P5 for encoding them, as well as the DECL application of the guidelines, and what benefits they have for the type of historically oriented, corpus searchable editions we are compiling. The examples will come from a digital edition of Trinity College Cambridge MS O.1.77, a pocket-sized late medieval medical handbook in Middle English and Latin, which I am editing for my PhD thesis.

Digital Editions for Corpus Linguistics

DECL is a project, based at the Research Unit for Variation, Contacts and Change (VARIENG) at the University of Helsinki, which aims at developing online editions that combine the accurate description of historical documents with the flexibility of search tools developed for linguistic computing. It was formed by three postgraduate students at VARIENG in 2007, who shared a dissatisfaction with extant tools and resources, and aimed to develop a more versatile and user-friendly model for digitised manuscripts of historical texts. The tools and framework are designed to meet the needs of small scale research projects and individual scholars. They are based on and compatible with version P5 of the TEI guidelines.

On the level of editorial principles, DECL editions adopt the opinion of Lass (2004) that digital editions should preserve the text as faithfully as possible, convey it in as flexible form as possible, and ensure that any editorial intervention remains visible and reversible, formulating it into three central principles of *Transparency*, *Flexibility* and *Expandability*. DECL editions aim to offer the user diplomatic transcriptions of the manuscripts into which linguistic, palaeographic and codicological features will be encoded. Additional layers of contextual, codicological and linguistic annotation can be added to the editions using standoff XML tagging.

Background

One of the most ubiquitous problems encountered in editing medieval manuscripts, is how to represent the numerous abbreviations in them. There is no established standard for encoding these abbreviations in digital format, and many digital editions still follow the practice inherited from traditional book editions of expanding them, either silently or in italics. From the point of view of historical linguistics this is somewhat problematic, especially in the light of some recent discussion over what is required of an edition or corpus in order to constitute reliable data (cf. i.a. Bailey 2004; Curzan and Palmer 2006; Dollinger 2004; Grund 2006). Most vocal in his criticism of existing practices has been Lass (2004), who demands that in order to serve as valid data for the historiography of language, a digital edition or a corpus should not contain any editorial intervention that results in substituting the scribal text with a modern equivalent.

Expanding abbreviations substitutes a symbol used by the scribe with a modern reading of it, which may, in the vast majority of cases, be obvious, and supported by research, but, by definition, also contains an element of editorial interpretation. In some cases this may have an impact on the data. For example, the irregularity of spelling of Middle English may result the editor to make decisions over which combination of letters a particular abbreviation stands for in text in which the abbreviated word may appear in several spelling variants.

TEI

The TEI P5 module for encoding glyphs and non-standard characters offers a few alternative ways of annotating them. The abbreviations may be annotated by `<g>` tag, indicating that they are glyphs, in which case they are defined by the gaiji module in the TEI header. Or the `<am>` and `<ex>` tags can be used to indicate the abbreviated sign and its editorial expansion. In these cases, the `<choice>` element may also be used to indicate that some of the elements are alternatives to each other. Or the whole word may simply be annotated as `<abbr>` to show that it contains abbreviations. The editor may also use the `<expan>` tag, indicating items which have been expanded without recording the abbreviation symbol.

The DECL guidelines uses an application of this that marks both the symbol, and its content, but does not require multiple elements inside a single word, as they can cause internal difficulties with stand-off tagging. We use the `<abbr>` element to mark that a word contains an abbreviation, the `<g>` element to tag the content of each abbreviation, and give the abbreviation symbol used for it as its attribute.

```
<abbr>su<g ref=”#crossed-p”>per
</g></abbr>
```

The expanded part of the abbreviation, which is in fact editors reconstruction and in some cases may up for debate gets enclosed in the `<g>` tags, and thus also marked as editorial - which is in accordance with the DECL principle of transparency.

Aims and Benefits

The XML code can be dynamically processed via XSLT transformation scripts to create documents which display either the abbreviations or expanded words according to the needs of the user, and DECL editions will also offer the user a customisable online interface, capable of displaying both. In addition to visual presentation and browsing, the interface will also offer corpus search and analysis functions, which can be extended to searches on the specified elements or attributes. Following the principles of open source and open access, they users of DECL editions will have full access to the code and may download and alter it, meaning that it is possible to alter the editorial decisions if the user is not satisfied with them.

In the poster I will give an illustrated presentation of how the process of encoding abbreviations progresses from manuscript images, via TEI XML code to its various forms of presentation.

References

Bailey, Richard W. (2004). The need for good texts: The case of Henry Machyn’s

Day Book, 1550–1563, *Studies in the history of the English language II: Unfolding conversations* (*Topics in English Linguistics* 45): 217–228.

Curzan, Anne and Palmer, Chris C. (2006). The importance of historical corpora, reliability, and reading, *Corpus-based Studies of Diachronic English*: 17–34.

DECL (Digital Editions for Corpus Linguistics). <http://www.helsinki.fi/varieng/domains/DECL.html>.

Dollinger, Stefan. (2004). ‘Philological computing’ vs. ‘philological outsourcing’ and the compilation of historical corpora: A Late Modern English test case, *Vienna*

English Working Papers (VIEWS), 13(2): 3–23.

Grund, Peter. (2006). Manuscripts as sources for linguistic research: A methodological case study based on the Mirror of Lights, *Journal of English Linguistics*, 34: 105–125.

Lass, Roger. (2004). 'Ut custodiant litteras: Editions, Corpora and Witnesshood', in: Marina Dossena and Roger Lass (eds.) *Methods and Data in English Historical Dialectology* (*Linguistic Insights* 16): 21–48.

TEI (Text Encoding Initiative). <http:/www.tei-c.org>.

VARIENG (Research Unit for Variation, Contacts and Change in English). <http://www.helsinki.fi/varieng/>

JGAAP 4.0 — A Revised Authorship Attribution Tool

Patrick Juola
Duquesne University
juola@mathcs.duq.edu

John Noecker, Jr.
Duquesne University
jnoecker@gmail.com

Mike Ryan
Duquesne University
michaelryan@acm.org

Sandy Speer
Duquesne University
speers@duq.edu

Authorship Attribution (Juola, 2006) can be defined as the inference of the author or her characteristics by examining documents produced by that person. For some time, we have been working on a system (JGAAP — Java Graphical Authorship Attribution Program) to use advanced statistics to perform this task while not demanding a high degree of expertise from the user (Juola, et al., 2008). With the recent release of JGAAP 3.2 and the near-term planned release of JGAAP 4.0, we are finally confident that we have a production quality system for general-purpose use.

We now report (and demonstrate) these recent improvements. JGAAP now incorporates nearly 20 different analytic methods (including eight different distance-based nearest-neighbor algorithms), more than 20 different event sets and models ranging from character- and word-based N-grams to reaction times, and several different preprocessors incorporating a wide variety of different document types including remote (Web-accessible) files and text extraction from different formats. We estimate that JGAAP is capable of performing more than 20,000 different types of analysis for authorship attribution or similar text classification tasks, with more being added as development continues.

Other improvements include:

- GUI improvements to enhance user-friendliness
- Enhanced graphical output capabilities

- Full report generation capacity for scholarly inspection of the results

- Creation of a command-line interface

- Automatic batch processing capacity for large-scale comparative testing

- Incorporation of the AAAC (Juola, 2004) test corpus into the demo for comparative testing purposes

- Dynamic loading of new methods to encourage new development

We are finally able to perform large-scale comparative analyses of different processing methods. We include here a short list of some JGAAP-related findings (published, submitted, or in preparation) :

- Introduction of a small number of character errors (as exemplified by modern OCR systems) does not substantially reduce accuracy with most methods.

- Symmetric ("commutative") distance-based methods tend to outperform asymmetric ones.

- Linear classifiers such as LDA tend to outperform nonlinear classifiers despite the apparent oversimplicity of the underlying model

- Character-based methods tend to outperform word-based ones for authorship attribution in Chinese

- Both cosine distance (normalized dot product) and simple event-based Kullback-Leibler divergence tend to be the best-performing methods for distance-based nearest-neighbor methods.

- The seminal word list of Mosteller and Wallace does not generally perform well for texts other than the Federalist Papers

Some of our findings have been submitted under separate cover to this conference, but we hope to present a summary of major results that have been achieved by June 2009 along with a demonstration of the newest version of the program. We also hope to provide examples of the sort of analysis that have been performed by JGAAP (and invite cooperation from interested researchers for further study).

Finally, we hope to demonstrate some example ad-hoc analyses during the session; it should be possible, for example, to demonstrate that "document length" or "words that are palindromes" do not perform well as Event/feature sets in less than ten minutes. While this is perhaps not interesting (no sensible person has proposed palindromes for authorship attribution), this clearly illustrates the ease-of-use and of result generation.

The Prioress and the Jew: Mining the Symbolic System through Lexical Genre Analysis of Modernizations

Nathan Kelber
University of Maryland, College Park
nkelber@umd.edu

My research uses quantitative analysis and reception theory in order to understand how early 19th century readers viewed and received the work of Chaucer. Using the HyperPo tool from the Text Analysis Portal (TAPoR), my analysis suggests that modernizations of "The Prioress's Tale" were more critically successful in the 19th century when they adulterated the original form of the tale beyond the threshold of the distinctiveness ratio (Hoover, 2008). I attribute this genre shift to a change in the symbolic systems between Christians and Jews which denigrated the legenda form in the 19th century. The historic adulteration of the legenda form along with the disappearance of its 'loci in life' illuminates both how and why current analysis has focused heavily upon the anti-Semitism of the tale while also coming to an impasse of critical scholarship.

Criticism of the anti-Semitic nature of the tale has tended to diverge and stagnate into what critic Lawrence Besserman calls 'hard' and 'soft' readings. Hard readings view both the tale and Chaucer as anti-Semitic, as was typical of 14th century England. Soft readings focus upon Chaucer's satire of the prioress, as well as adducing ambivalent historical evidence of the relationship between Jews and Christians, in an attempt to redeem Chaucer, the 'Father of English.' My research shifts this dualism by suggesting that contemporaries of Chaucer saw the construction of the Jew, or virtual Jew as posited in the work of Sylvia Tomasch, as a purely symbolic construct inherent within the form and social function of the legenda. My analysis of the degradation of the legenda is framed by the reception theory of Hans Robert Jauss.

Jauss's theory, elucidated in Toward an Aesthetic of Reception, diachronically links literary studies and literary history by tracing the historical shift of genres through changes to readers' 'horizon of expectations.' According to Jauss, texts propagate through their ability to conform to and disturb the expectations of their readers. When texts are well-received, they exhibit congruency with the history and sociology of their audience. The reception of deviations from the horizon of expectations signals the historical and social approval or denigration of a text's ability to represent the norms and realities of its readers. The structure of genre then is more than mere literary fashion. The diachronic change of genre reception reflects instabilities in the specific social and historical norms and realities of readers at the time. Stable genres reflect social and historical systems that are well-established. Changing genres reflect periods of social and historical change.

Within the 14th century context of 'The Canterbury Tales,' the legenda of 'The Prioress's Tale' is a 'culinary art' (Unterhaltungskunst) because it demands no horizontal change of expectations (Jauss, 1982). The legenda represents an older form of art designed for what Margaret Hallisy calls 'simple believers.' The characters function as simple allegories of weakness, holiness, and evil. The social function of the legenda is to show how the holy but weak may paradoxically triumph over a powerful evil with the help of God. The strength of this simple allegorical form in subsequent modernizations of the tale can be tracked through text analysis software such as HyperPo.

The form of the legenda is reliant upon clear designations of good, evil, and weakness. The saint must always be portrayed as both good and weak. The villain is purely evil with no sense of redemption. I have used HyperPo to examine a lexical field centered about these terms in Robert Lipscomb's 1795 edition, William Wordsworth's 1820 edition, and Robert Anderson's Middle English version in The Works of the British Poets. I have chosen Anderson because he is the likely source for both writers.

My data from HyperPo suggests that Wordsworth's text is a fairly good modernization. He retains the rhyme royal form, uses similar syntax, and foregoes translation where he feels the modern word does not capture the essence of the original Middle English. The lexical field of the legenda is a strong match with Anderon (92% raw match), but shows some weakening. Concerning reception, reviews of Wordsworth's text in 1820 either panned or ignored the modernization. This seems to suggest a disconnect between the function of the legenda form and Wordsworth's 19th century audience.

Lipscomb's text is not a close modernization. Analysis shows that the lexical field of the legenda is debased to the point that the text only weakly resembles the form (62% raw match). This match is below the Distinctiveness Ratio of .67 which David L. Hoover advocates as warranting attention (2008). It might be said that Lipscomb's text has ceased to be a legenda at all. Reception of Lipscomb's text was very positive which shows that his changes create a text that is more congruent with the aesthetics of contemporary readers. In comparison to Wordsworth and Anderson, Lipscomb's text was a watered-down Chaucer which Lipscomb "pruned of indelicacies."

HyperPo, as a text analysis tool, is an effective way of creating and organizing lexical data. The addition of raw counts, relative weights, and Z-scores gives the critic a variety of measurements for their data. Working between Middle English and Modern English creates problems. Anderson's nonstandard spelling causes inaccurate word counts where some words (moder/modre/mother) are counted separately and other words are counted together ("sone" meaning both "son" and "soon"). Critics must be careful if filtering common words that they do not remove words which might affect their data.

My research suggests that Wordsworth's modernization was a fairly accurate translation of Chaucer, but this is also the reason it was a failure. Contemporary readers saw Chaucer and Middle English as barbaric and crude. Lipscomb's version, on the other hand, softened the language and delivered a less vitriolic modernization. Lipscomb worked within the framework of Dryden's Fables Ancient and Modern, published in 1700, which drastically changed lines, added material, changed rhyme schemes, and censored what he considered indecent. Dryden's modernizations were very popular and fit more closely into the horizon of expectations of the time. The social function of the legenda form, if distant in Chaucer's time, was no longer apparent to the 19th century reader. In turn, changes within the symbolic system between Christians and Jews render the Prioress's character opaque to modern readers. Her anti-Semitism is troubling because it confronts us in the aftermath of a long history of Jewish persecution which culminated in the recent history of the holocaust.

Bibliography

Anderson, Robert, ed. (1795). The Prioress's Tale. The Works of the British Poets. London.

Benson, Larry, ed. (1987). The Riverside Chaucer. Boston: Houghton Mifflin.

Graver, Bruce. (1998). Translations of Chaucer and Virgil. Ithaca: Cornell UP.

Hallissy, Margaret. (1995). A Companion to Chaucer's Canterbury Tales. Westport: Greenwood.

Hoover, David L. (2008). Quantitative Analysis and Literary Studies. A Companion to Digital Literary Studies, ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell.

HyperPo. (2008). Text Analysis Portal For Research. http://portal.tapor.ca/portal/portal (accessed 20 October 2008).

Jauss, Hans Robert. (1982). Trans. Timothy Bahti. Toward an Aesthetic of Reception. Minneapolis: U of Minn P, 1982.

Lipscomb, William, ed. (1795). The Prioress's Tale. The Canterbury Tales, complete in a Modern Version. Oxford.

Tomasch, Sylvia. (2000). Postcolonial Chaucer and the Virtual Jew. The Postcolonial Middle Ages. Ed. Jeffrey Jerome Cohen. New York: Palgrave, pp. 243-260.

Tyrwhitt, Thomas, ed. (1795). The Canterbury tales of Chaucer.

Wilsbacher, Greg. (2005).Lumiansky's Paradox: Ethics, Aesthetics and Chaucer's 'Prioress's Tale.' College Literature. Vol. 32, Iss. 4: 1-29.

Wordsworth, William. (1947). The Prioress' Tale. Wordsworth's Poetical Works. Ed. E. Selincourt. London: Oxford UP.

Wu, Duncan. (1993). Wordsworth's Reading 1770-1799. Cambridge: Cambridge UP.

An Approach to Information Access and Knowledge Discovery from Historical Documents

Fuminori Kimura
Ritsumeikan University
fkimura@is.ritsumei.ac.jp

Akira Maeda
Ritsumeikan University
amaeda@media.ritsumei.ac.jp

1. Introduction

Recently, libraries, governments and major internet providers, are forming consortiums to preserve historical documents stored in libraries. (e.g. Google Book Search, Open Content Alliance, World Digital Library, Hathi Trust, etc.). It means that more and more old text contents will be accessible on the internet in the near future. Obviously, huge amount of knowledge in old documents is as important as recent born-digital documents typically available on the web, because old documents are the collection of wisdom from B.C. Thus, it might be useful to be able to access such old documents. Moreover, it is very useful to discover hidden knowledge and wisdom written in these old documents.

In order to realize this purpose, it is necessary to retrieve important information from old documents. However, it is not always easy to retrieve old documents, mainly due to the substantial change in language and culture over time. Therefore, we need a method to access old documents written in ancient language using a query in modern language. We call this method "Cross-Age Information Retrieval". Moreover, we should consider the cultural difference over time, even for the same language. For this, we need a method of "Cross-Cultural Information Retrieval".

Most of the research on information retrieval and information access focus on documents written in modern language, but we believe that knowledge and wisdom written in old documents provide rich and valuable information which are not available in modern language documents, especially in web contents.

We propose a "Cross-Age Information Retrieval" method in order to tackle these problems. It aims to discover hidden knowledge and wisdom written in old documents.

2. Related Work

Much research on Cross-Language Information Retrieval has been conducted in the last 10 years, with the background of the rapid growth of the web around the world since the middle of 1990's. Various approaches, including query translation, document translation, and the use of intermediate language has been studied, and for certain language pairs (e.g. between European languages), adequate retrieval effectiveness has been achieved.

On the contrary, there is very little research on information retrieval method for historical documents, and most of which are based on simple keyword matching. Recently, some approaches have been proposed to access historical documents, and it could be regarded as a kind of Cross-Age Information Retrieval (Gerlach et al. 2007; Khaltarkhuu et al. 2006). Our goal is to establish a more effective and sophisticated retrieval method that considers not only language difference over time, but also cultural difference between languages and ages.

3. The Proposed Method

We adopt dictionary-based query translation approach, since it is proven to be the most effective method for Cross-Language Information Retrieval. In order for dictionary-based methods to be effective, we need to use precise and comprehensive dictionaries for both modern language and ancient language. From these two dictionaries, we try to discover relationships between entries in those dictionaries, and to "translate" the query terms in modern language into equivalent terms in ancient language. For this translation process, we propose the following method (Fig. 1):

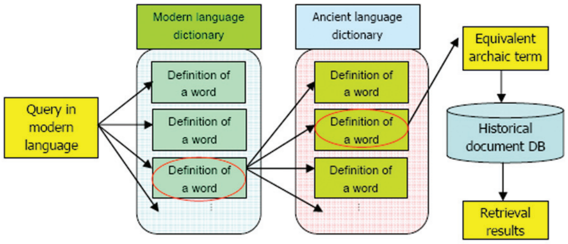


Fig. 1 Overview of the proposed method for Cross-Age Information Retrieval.

1. For each entry in the modern dictionary, we look for an equivalent entry in archaic word dictionary by calculating the similarities between the definition of the modern word and all the definitions of the archaic words. For this process, we can use standard text similarity measure based on vector space model

and tf-idf term weighting scheme.

2. Then, we take the most similar definition in archaic word dictionary, and that entry (headword) is regarded as an equivalent of the modern word.

3. If more than one equivalent entry exists, we disambiguate the translation candidates using the term association measure such as mutual information, to find the most equivalent archaic word for the modern language word.

4. Document Collections

Currently, it is not very easy to obtain historical documents in text format. However, some digital libraries (e.g. Google Book Search, Open Content Alliance, etc.) are ready to provide their collection of historical documents in text format for research purposes. Moreover, there are numerous existing old documents available online. In Japan, there is a volunteer-based effort called "Aozora Bunko" to digitize and to make accessible over 7,000 copyright-expired classic literatures online. Also, many universities and institutions have already been providing collections of old documents in text format. We can use these huge collections of old documents for our proposed method.

For now, we are focusing on a Japanese historical document called "Hyohanki", which was written in late Heian era (12th century) in Japan. It is a valuable resource for the research of Japanese culture of that time period. An example of its original copy is shown in Fig. 2. Although some part of it has been deteriorated and missing, all of the existing pages are digitized into text format. The existing pages consist of 2,488 diary entries.

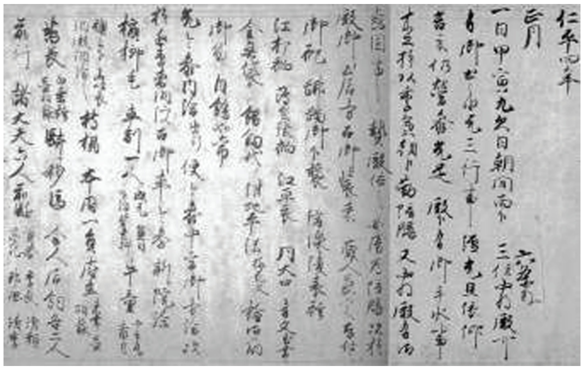


Fig. 2 Example of the original copy of a historical Japanese document "Hyohanki".

5. Language Resources

As described in Section 3, we need dictionaries in order to translate modern language query into archaic term(s). In the case of "Hyohanki", we can use some existing electronic dictionaries available in CD-ROM. For Japanese modern language, we use "Kojien", one of the most famous and comprehensive Japanese language dictionaries. For ancient language, we use "Kokugo-Daijiten", which covers not only modern words but also archaic words.

6. Preliminary Experiment

We conducted a preliminary experiment to test the precision of "Cross-Age retrieval" by our proposed method. In this experiment, we used diary entries of "Hyohanki" as the ancient Japanese document collection, and prepared 3 modern Japanese queries, "戦争 (war)", "法要 (Buddhist service)", and "裸足 (bare foot)". Since each query has an equivalent archaic term in different wording, no relevant documents can be retrieved if we use these modern term queries. Note that, we consider one diary entry as one document.

modern Japanese query	translated terms	relevant / retrieved
戦争 (war)	軍, 戦	10 / 37 (27%)
法要 (Buddhist service)	仏事	109 / 110 (99%)
裸足 (bare foot)	跣, 裸足, 跣足	27 / 27 (100%)
死亡 (death)	没	2 / 13 (15%)

Table 1 shows the original modern Japanese query, ancient Japanese term(s) translated by the proposed method, and the precision of retrieval using the translated term(s). For the queries "法要 (Buddhist service)" and "裸足 (bare foot)", the proposed method worked quite well and chieved almost 100% precision (the ratio of relevant documents in retrieved documents). However, the query "戦争 (war)" resulted in very poor precision (27%). The reason for it is that the proposed method returned two translation candidates (i.e. "戦" and "軍") for this query. If we take only "戦" as the translated query, we could achieve 100% precision, but if we take only "軍", we could obtain only 3.6% precision. It is because the archaic term "軍" has not only a meaning "war", but also other meanings like "general (officer)" and "army". The query "死亡 (death)" also resulted in very poor precision (15%) and the reason for it is that the translation "没" has several meanings, "death", "deprivation" and "sunset". These results suggest that we could improve the precision if we incorporate a suitable disambiguation method for the translated archaic terms. For that purpose, we could apply existing disambiguation methods used in Cross-Language Information Retrieval, such as mutual information, etc.

7. Conclusion

In this paper, we proposed a novel information retrieval technique called "Cross-Age Information Retrieval", which can be used to access old documents written in ancient language using a query in modern language. We conducted a preliminary experiment to test the precision of cross-age retrieval by our proposed method. The experimental results showed that our proposed method is potentially useful for cross-age retrieval. Although our proposed technique is still in an early stage, we believe that we can achieve adequate retrieval effectiveness by incorporating techniques used for Cross-Language Information Retrieval.

Our goal is not only to realize cross-age retrieval, but also to extend this technique to more advanced text mining applications in order to discover hidden knowledge and wisdom from large amount of premodern documents which are now available in digital form.

Our future work includes resolving ambiguity of translated archaic terms, large-scale experiments in other languages such as English, consideration of cultural difference over time, and thus extending our technique to realize cross-age, cross-cultural, and cross-language information access.

References

Gerlach, A. E. and Fuhr, N. (2007). Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2007)*, pp. 333-341, 2007.

Khaltarkhuu, G. and Maeda, A. (2006). Retrieval Technique with the Modern Mongolian Query on Traditional Mongolian Text. In *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL2006)*, pp. 478-481, 2006.

Modulating Style (and Expectations): An Experiment with Narrative Voice in Faulkner's *The Sound and the Fury*

Caitlin Crandell
Stanford University
crandell@stanford.edu

Emily Gong
Stanford University
mlygng@stanford.edu

Rachel Kraus
Stanford University
rkraus1@stanford.edu

Tiffany Lieu
Stanford University

Jacob Mason-Marshall
Stanford University
jacobm2@stanford.edu

Faulkner's *The Sound and the Fury* provides an interesting test bed for stylistic analysis of narrative and authorial voice. Readers of the text understand that the work contains at least four different "voices," and if we include the Appendix that Faulkner wrote after the book's initial publication, then we have five distinct voices, namely those of Benjy, Quentin, Jason, the Omniscient narrator of the fourth section, and then the somewhat mysterious voice of the Appendix that some have called Faulkner's own voice. As part of Dr. Mathew Jockers's humanities computing seminar at Stanford this fall, the authors of this proposal conceived of an experiment to investigate narrative voice and utilize authorship attribution techniques to explore the extent to which Faulkner is able to "create" distinct narrative styles within *The Sound and the Fury*.

Starting with a traditional reading of Faulkner's novel, our group employed the traditional tools of literary analysis as a sort of "control." We explored the differing styles content, and structure of the five separate narrative voices in *The Sound and the Fury*. We paid close attention to the stylistic changes, and we hypothesized about what qualities separate the different sections. We put all of these observations into an aggregated document; a traditional "reading" of style in Faulkner's text.

We then began a phase of speculation based upon our understanding of the sort of stylistic data that can be generated by a computer algorithm. We speculated about how or if these qualities identified in our traditional reading might be detectable via computer-based text analysis. We then drew up a series of hypotheses and predictions about what sort of differences might be made evident by a quantitative, computer analysis of the text.

Despite the significant shifts in style detected by readers, our first prediction was that in terms of the most common words, Faulkner's narrative would be fairly consistent. Our prediction was informed by current studies in authorship attribution, which suggest that, with frequently occurring words at least, differences between authors are greater than difference between works by the same author. With *The Sound and the Fury*, we had a particularly compelling test case since Faulkner worked very hard to create unique narrative perspectives, even changing between first and third person.

Our second prediction involved the very noticeable changes a reader detects when moving from section to section. These changes constitute a sort of "narrative dissonance" that repeatedly "jars" the reader as he moves between sections. We speculated that these effects would be difficult for a computer to detect.

We formulated these and other predictions/hypotheses into a secondary document, and then began the more practical work of developing the tools necessary to test our hypotheses. This work involved leveraging existing tools such as Patrick Joula's JGAAP and also creating new tools. Most challenging has been our effort to create a "dissonance" detector, a tool which at the time of this writing is showing great promise but still under active development.

The paper we propose here represents a sort of experiment in which we are both the subjects and the investigators. While we are overtly exploring narrative voice in Faulkner we are simultaneously investigating the role that a computer-based methodology might play in a conventional study of literature. On this point we are inspired by Steve Ramsay's observation that texts are "seldom . . .transformed algorithmically as a means of gaining entry to the deliberately and self-consciously subjective act of critical interpretation." In our final analysis, we transform our text into a new sort of literary artifact, a statistical matrix that allows us to work in a reverse order—starting with objective, quantified facts and moving to subjective interpretations of those data. Ultimately, we compare our objective and subjective methods and their respective conclusions to determine to not only explore the formal aspects of an author's narrative style, but to imagine what weight quantified data can bring to the traditional literary enterprise.

Les techniques informatiques au service des connaissances musicales de la renaissance

Florence Le Priol
Université Paris-Sorbonne
Florence.Le_Priol@paris-sorbonne.fr

Cristina Diego Pacheco
Université Nancy 2
cristina.diego-pacheco@univ-nancy2.fr

Louis Jambou
Université Paris-Sorbonne
Louis.Jambou@paris-sorbonne.fr

Le travail présenté dans cet article est la concrétisation d'une collaboration interdisciplinaire entre le groupe « Lexique Musical de la Renaissance (LMR) » et l'équipe « Langages Logiques Informatique et Cognition (LaLIC) » visant à construire un outil pour un nouveau savoir musical. Comme nous l'avions présenté à Digital Humanities 2007, l'objet du projet est le développement de cet outil qui s'appuie sur une base de données relationnelle et une interface web multilingue (français, anglais, espagnol et portugais). On y présente le langage musical depuis la naissance de la théorisation musicale en langue vernaculaire jusqu'au seuil de la formation du langage tonal. Aujourd'hui opérationnel (http://www.pm.paris4.sorbonne.fr/LMR/), cet outil propose :

- un lexique musical de la renaissance,
- un dictionnaire du langage musical,
- un ensemble de traités musicaux en langue espagnole.

Lexique musical de la renaissance

Le lexique musical de la renaissance (LMR) est un lexique multilingue, cumulatif. Chaque terme regroupe, en chaque langue mais dans la succession alphabétique intégrale, toutes les citations réunies par les collaborateurs. La constitution de ce lexique exige dans un premier temps un travail de lecture musicologique et linguistique des traités et une sélection des citations les plus pertinentes en rapport avec l'acte de la production et la pensée musicales. Dans un deuxième temps, le travail de saisie dans la base de données est réalisé grâce aux interfaces de saisie mises à disposition. Pour chaque traité, les citations sélectionnées et les termes du lexique qui y sont attachées sont saisies.

Actuellement, ce lexique, constitué à partir de l'étude de plus de 80 traités en espagnol (73) et en français (8), comprend près de 5500 citations (4223 en espagnol et 1258 en français) et plus de 5100 termes (3822 en espagnol et 1327 en français).

Dictionnaire du langage musical

Le dictionnaire du langage musical, issu de la réflexion musicologique et linguistique, comprend un répertoire raisonné et sélectif de citations pertinentes des différents langues techniques musicales vernaculaires. La sélection, dans la langue considérée, est opérée par un regroupement de citations sous des champs sémantiques ou acceptions qui donnent lieu à des définitions. Le nombre de champs sémantiques est une conséquence du contenu signifiant des citations elles-mêmes. Le dictionnaire se présente d'une part sous une forme unilingue où chaque langue est traitée indépendamment, d'autre part sous une forme multilingue où, sous des termes ou entrées multilingues des citations sont regroupées sous une même acception. Les termes/entrées multilingues ne résultent pas d'une traduction directe ou littérale mais du signifiant même des citations.

Traités musicaux en langue espagnole (TME)

Le TME est une base de données qui regroupe un corpus de textes musicaux de la Renaissance écrits en langue espagnole. Il vient ainsi compléter des projets semblables proposés par le TML, le LmL, le SMI et CANTO. Le TME vient compléter le corpus saisi dans le LMR en offrant aux chercheurs des citations contextualisées dans les sources primaires, manuscrites ou imprimées. Il permet d'enrichir les définitions lexicographiques à travers une pensée musicale plus large et d'intégrer les citations du LMR dans leur contexte d'origine. Chaque citation du LMR est, en effet, reli√©e, par un simple « clic » au traité en texte intégral dont elle est issue. Trois traités sont disponibles, les autres sont en cours de saisie.

Mise en oeuvre informatique

Les techniques informatiques utilisées pour mettre en place cet outil, tant pour la zone de saisie (accès sécurisé) que pour la zone de consultation (accès libre), sont celles utilisées pour développer des applications disponibles en ligne : APACHE, PHP, JavaScript, MySQL, XML, XSLT et CSS.

La base de données relationnelle, implémentée avec MySQL, est l'élément principal du projet, où sont capitalisées toutes les données musicologiques et linguistiques provenant des traités. Elle est organisée en deux parties, d'un coté une base intermédiaire où sont enregistrées toutes les données saisies, de l'autre, la base de données définitive où les données intermédiaires sont transférées après validation. La base définitive est organisée en plusieurs tables liées afin de faciliter l'accès aux informations, par exemple, la table « tblauteur » regroupe les informations sur les auteurs des traités, la table « tblsource » comprend les informations bibliographiques des traités, les termes du lexique sont enregistrés dans la table « tblentree » ...

Deux types d'interfaces ont été développées : les interfaces pour la saisie des données et celles pour la consultation. Ces interfaces sont développées en PHP et JavaScript afin d'offrir une compatibilité multi-plateforme.

Les interfaces de saisies (par exemple la figure 1), dont l'accès est sécurisé par mot de passe, sont constituées de formulaire permettant aux chercheurs de saisir aisément leur travail, sans avoir à accéder directement à la base de données. Ces formulaires ont été élaborés afin de suivre au plus près la réflexion musicologique et linguistique.

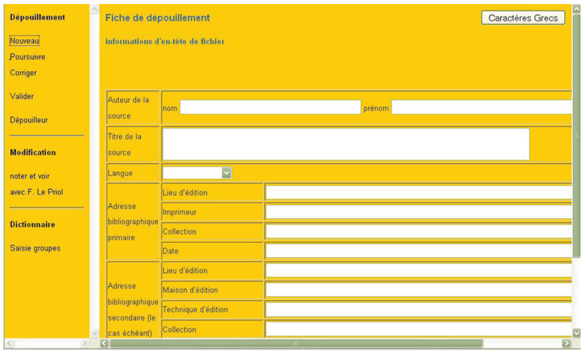
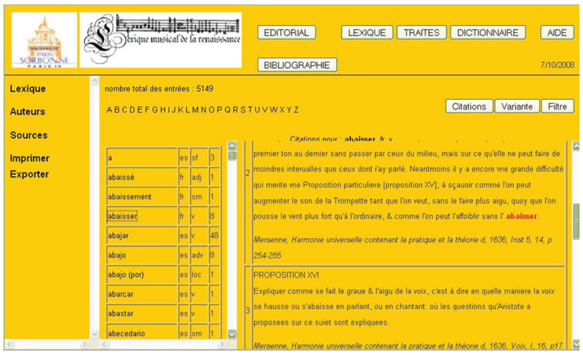


Figure 1. Premier formulaire de saisie pour un nouveau traité

La consultation des données est en accès libre, l'interface est multilingue : français, anglais, espagnol et portugais. Elle donne accès au LMR (figure 2), au dictionnaire et au TME (figure 3).

Le TME est réalisé en utilisant XML, XSLT et CSS. La standardisation des traités au format XML et l'utilisation des feuilles de styles XSLT et CSS permettent d'avoir une homogénéité dans la présentation et de lier le LMR au TME. En effet, lors de la consultation du LMR, un simple « clic » sur l'icône situé après la citation, affiche le texte du traité dans lequel la citation est surlignée comme dans l'exemple de la figure 4.



7/10/2008

nombre total des entrées : 5149

Citations pour : abaisser fr v

a	es	sf	3
abaissé	fr	adj	1
abaissement	fr	sm	1
abaisser	fr	v	8
abajar	es	v	48
abajo	es	adv	8
abajo (por)	es	loc	1
abarcar	es	v	1
abastar	es	v	1
abecedario	es	sm	1

2 premier ton au dernier sans passer par ceux du milieu, mais sur ce qu'elle ne peut faire de moindres interualles que ceux dont i'ay parlé. Neantmoins il y a encore vne grande difficulté qui merite vne Proposition particuliere [proposition XV], à sçauoir comme l'on peut augmenter le son de la Trompette tant que l'on veut, sans le faire plus aigu, quoy que l'on pousse le vent plus fort qu'à l'ordinaire, & comme l'on peut l'affoiblir sans l' **abaisser**.

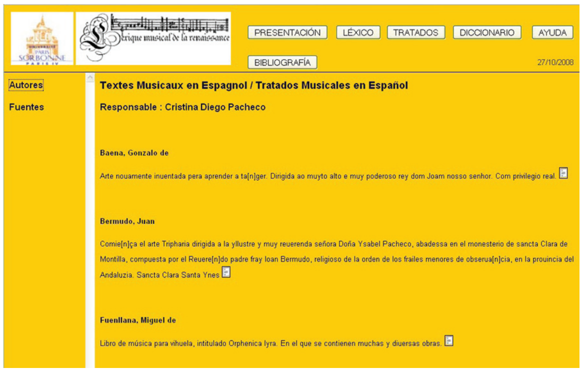
Mersenne, Harmonie universelle contenant la pratique et la théorie d, 1636, Inst 5, 14, p 254-255

3 PROPOSITION XVI

Expliquer comme se fait le graue & l'aigu de la voix, c'est à dire en quelle maniere la voix se hausse ou s'abaisse en parlant, ou en chantant: où les questions qu'Aristote a proposees sur ce suiet sont expliquees.

Mersenne, Harmonie universelle contenant la pratique et la théorie d, 1636, Voix, I, 16, p17

Figure 2. Interface (en français) de consultation du LMR



27/10/2008

Textes Musicaux en Espagnol / Tratados Musicales en Español

Responsable : Cristina Diego Pacheco

Baena, Gonzalo de

Arte nouamente inuentada pera aprender a ta[n]ger. Dirigida ao muyto alto e muy poderoso rey dom Joam nosso senhor. Com privilegio real.

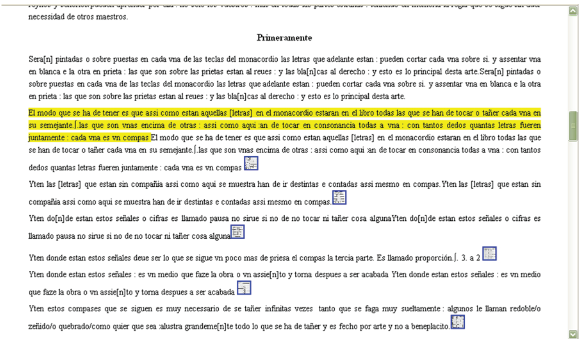
Bermudo, Juan

Comie[n]ça el arte Tripharia dirigida a la yllustre y muy reuerenda señora Doña Ysabel Pacheco, abadessa en el monesterio de sancta Clara de Montilla, compuesta por el Reuere[n]do padre fray Ioan Bermudo, religioso de la orden de los frailes menores de obserua[n]cia, en la prouincia del Andaluzia. Sancta Clara Santa Ynes

Fuenllana, Miguel de

Libro de música para vihuela, intitulado Orphenica lyra. En el que se contienen muchas y diuersas obras.

Figure 3. Interface (en espagnol) d'accès au TME



[illegible] necessidad de otros maestros.

Primeramente

Sera[n] pintadas o sobre puestas en cada vna de las teclas del monacordio las letras que adelante estan : pueden cortar cada vna sobre si. y assentar vna en blanca e la otra en prieta : las que son sobre las prietas estan al reues : y las bla[n]cas al derecho : y esto es lo principal desta arte.Sera[n] pintadas o sobre puestas en cada vna de las teclas del monacordio las letras que adelante estan : pueden cortar cada vna sobre si. y assentar vna en blanca e la otra en prieta : las que son sobre las prietas estan al reues : y las bla[n]cas al derecho : y esto es lo principal desta arte.

El modo que se ha de tener es que assi como estan aquellas [letras] en el monacordio estaran en el libro todas las que se han de tocar o tañer cada vna en su semejante.ʃ.las que son vnas encima de otras : assi como aqui :an de tocar en consonancia todas a vna : con tantos dedos quantas letras fueren juntamente : cada vna es vn compas El modo que se ha de tener es que assi como estan aquellas [letras] en el monacordio estaran en el libro todas las que se han de tocar o tañer cada vna en su semejante.ʃ.las que son vnas encima de otras : assi como aqui :an de tocar en consonancia todas a vna : con tantos dedos quantas letras fueren juntamente : cada vna es vn compas

Yten las [letras] que estan sin compañia assi como aqui se muestra han de ir destintas e contadas assi mesmo en compas.Yten las [letras] que estan sin compañia assi como aqui se muestra han de ir destintas e contadas assi mesmo en compas.

Yten do[n]de estan estos señales o cifras es llamado pausa no sirue si no de no tocar ni tañer cosa algunaYten do[n]de estan estos señales o cifras es llamado pausa no sirue si no de no tocar ni tañer cosa alguna

Yten donde estan estos señales deue ser lo que se sigue vn poco mas de priesa el compas la tercia parte. Es llamado proporción.ʃ. 3. a 2

Yten donde estan estos señales : es vn medio que faze la obra o vn assie[n]to y torna despues a ser acabada Yten donde estan estos señales : es vn medio que faze la obra o vn assie[n]to y torna despues a ser acabada

Yten estos compases que se siguen es muy necessario de se tañer infinitas vezes tanto que se faga muy sueltamente : algunos le llaman redoble/o zeñido/o quebrado/como quier que sea :alustra grandeme[n]te todo lo que se ha de tañer y es fecho por arte y no a beneplacito.

Figure 4. Extrait du traité de G. Baena

Références récentes

Journée d'études « Musique ancienne en Sorbonne », 2 juin 2006, Maison de la Recherche, Sorbonne, Paris. Présentation par Florence Le Priol et Louis Jambou des *Problmatiques et enjeux du « Lexique Musical de la Renaissance »* (LMR)

Digital Humanities 2007, University of Illinois, Urbana-Champaign, 4-7 juin 2007, Louis Jambou et Florence Le Priol, *« Un outil pour un nouveau savoir musical »*, p 98-100

XI Settimana di alti studi rinascimentali, Invisibili Fili, Musica, Lessicografia, Editoria e tecnologie dell'informazione tra XVIe XXI secolo, Università degli Studi di Ferrara, 28-30 May 2009

Du lexique au dictionnaire musicologique :

- Aspects Théoriques (Louis Jambou)
- Aspects informatiques (Florence Le Priol)

AfricaMap Release I, Beta

An Infrastructure for Collaboration

http://africamap.harvard.edu

Benjamin Lewis
Harvard University
blewis@cga.harvard.edu

Suzanne Blier
Harvard University
blier@fas.harvard.edu

Peter Bol
Harvard University
pkbol@fas.harvard.edu

In November of 2008 the Phase I release (beta) version of AfricaMap was made available to the Harvard community and the public. The application can be accessed at http://africamap.harvard.edu.

AfricaMap sets out to address the problem of data availability for Africa. Much public data exists, but it is so difficult to discover, let alone obtain that many research projects on Africa spend much of their budget gathering data. Most people in Africa have an even harder time accessing mapping of their own territories. When researchers do gather data it is often once again lost because there is no place to store it where it can be found.

The AfricaMap project represents a framework for organizing Africa data which can also be applied to other parts of the world. At its core is a digital base map of the continent, viewable dynamically at a range of scales, and composed of the best cartographic mapping available. Behind the scenes a gazetteer starting with over 1 million place names provides rapid navigation to specific locations on a vast landscape. As more detailed mapping becomes available it will be added to the system. There is no limit in terms of hardware or software to the amount of data that can be added to the system.

AfricaMap is not be tied to a certain discipline but is interested in storing or referencing data from all disciplines. AfricaMap will encourage collaboration. Researchers will be able to define geographic areas of research so that others can find out about their work. The system employs a Services Oriented Architecture (SOA), which means that all the data that the system displays does not have to be stored on AfricaMap's servers. The data that is stored on the AfricaMap servers is made available to other applications as map services

The idea for AfricaMap was developed under a Provost Funds for Innovative Technology funded project that is now being overseen jointly by Suzanne Blier and faculty and staff at the Harvard Center for Geographic Analysis (Peter Bol, Wendy Guan, Ben Lewis). It has the dual aim of supporting Harvard research that involves GIS work on the continent and of making data created in the course of research available to others.

AfricaMap System Characteristics:

Web-based – The system takes advantage of the latest techniques for making large amounts of data and mapping discoverable and usable through a standard web browser.

Public access to holdings – Core holdings will be put in the public domain or licensed using a Creative Commons type license wherever possible. This means that researchers anywhere in the world will be able to download and use these original materials without major restriction.

Encourage replication – One reason Africa data is hard to find is that the data which exists is not yet well replicated. By contrast the base map for the United States (the Digital Raster Graphics files) area easy to find and exist on hundreds of servers.

Base mapping – Historic base maps for Africa are developed by scanning, cropping, and georeferencing maps from the Harvard Map Collection and elsewhere. Maps are digitized at a range of scales and for a range of time periods.

Dynamic gazetteer – The gazetteer together with the base map form the core of the AfricaMap system. These two datasets support one another over time, allowing the gazetteer to grow and improve, which will make it easier to find places on the base map.

Collaborative approach – Some tools to support collaboration between researchers are provided. In the first version a permalink feature will allow any view of the system to be captured in a URL which can be shared. In the next phase user created maps and map markup tools are anticipated. Researchers will be able to download base mapping and other datasets.

Multiple scales – The system will support research at a variety of scales from sites or cities to country or continent-wide projects.

Multiple media types – The system will support access to many types of media in addition to spatial data, including photos, maps, text, video, audio, and KML for Google Earth display.

Long term data access – Once maps are scanned, digitized, georeferenced it should not be necessary for anyone in the world to repeat that work. Making digital materials available over time is not easy because technology changes. Techniques will be used to ensure long term access to public domain digital materials wherever possible.

Improves over time – While the Harvard Map Library has large Africa holdings, it does not always have all maps for a given series, and there may be important series which it does not have. The goal is to fill in holes in the collection over time by sharing with other libraries and collections. Users will be able to submit data to Harvard using an online form.

Usability – Ease of use is of primary importance. It must be easy and quick for non-technical people to find the information they need. Researchers are the end users of this system and will be consulted frequently to guide the design of the user interface.

Text-based search of contents – Google-type text search against the contents of the entire system is possible with results displayed on the map.

Interdisciplinary approach –- The system will bring together mapped data (and facilitate the mapping of data) from a wide range of disciplines including archaeology, public health, history, linguistics, literature, zoology, natural resources to name a few.

Global approach – The goal is to create a technical framework to support research on Africa which could also be applied to other parts of the world. It is hope that aspects of AfricaMap will be useful for organizations based in Africa whether it is the underlying data, data hosting services, map services, or the AfricaMap software.

Scalable – The data in the system will be cached as it is used. This approach greatly increases performance and reduces server load, making the system far more scalable than a traditional web-GIS.

Services oriented architecture (SOA) – The system will support access by other web and desktop systems and will in turn be able to access and display the maps on AfricaMap directly via web services. This means that other systems will not have to download the data to access it within their applications.

Cross Platform – AfricaMap can serve data services to other types of GIS platforms including ArcMap desktop and ArcGIS Server. In addition, AfricaMap can display data served up from other platforms. Data formats used will be open specification ones such as GeoTIFF, JPG2000, KML, and Shape.

Open Source – The software that runs AfricaMap is Open Source and available for users and organizations inside and outside Harvard to obtain and build upon.

Forging the Future: New Tools for Variable Media Art Preservation

Marilyn Lutz
University of Maine
lutz@maine.edu

Jon Ippolito
University of Maine
jippolitto@maine.edu

Sharon Quinn Fitzgerald
University of Maine
quinn@maine.edu

Richard Rinehart
University of California, Berkeley
Rinehart@berkeley.edu

While the number of tools for making and distributing culture has exploded in the last half-century, it is hard to find a tool for preserving inherently ephemeral media, such as performances, installations, and digital artifacts that have been created using digital, biological, performative and other variable media. The increased use by artists of multi-media, digital and internet media raises questions about conventional strategies by which society preserves, cares for, and redisplays these cultural artifacts. While the most obvious vulnerability of new media art is rapid technological obsolescence, the study of its other aspects that defy traditional conservation—including hybrid, contextual, or 'live' qualities—has provoked an investigation into new strategies for ephemeral media art preservation. While we can preserve something of new media art, it will inevitably be changed. Forging the Future, a consortium of museums and cultural heritage organizations, aims to fill the gap with a free, practical toolset designed to rescue variable media works from oblivion by understanding possible alterations, distilling which matter most to the artist, and planning accordingly.

The Forging the Future tools are based on research into innovative preservation standards and strategies by its members as part of the Variable Media Network (VMN) and Archiving the Avant-Garde working group. These include the Franklin Furnace Database, which catalogs past versions of a work and variable media artworks and events contained in small to midsize collections; the Digital Assets Management Database, which manages digital metadata that is directly relational to all the tools; the Variable Media Questionnaire (VMQ), which contains data and metadata necessary to migrate, re-create, and preserve cataloged variable media objects: What is the work in its original incarnation? What could the work be in later incarnations? Beta versions of these tools are in use by institutions ranging from major museums such as the Whitney Museum of American Art to small alternative spaces such as New Langton Arts.

Scheduled for a public release in Spring 2009, the Forging the Future toolset will include features that help the separate tools interconnect with each other and with potential databases in museums, libraries, and archives. In addition to the Franklin Furnace database and the third generation VMQ web service, these include the VocabWiki, a multi-institutional effort to define key terms applicable to new media and variable media artifacts; Media Art Notation System (MANS), an XML specification, based on MPEG-21 for translating data and relationships among tools; and the Forging the Future Metadata Registry, which enables users to track the same creator or work across different collections.

This poster presents the Forging the Future interconnected tool set, and focuses specifically on the VMQ and its integration with MANS, complimented by the VocabWiki and the Metadata Registry. The VMQ is an instrument for determining creators' intent as to how their work should be categorized, seen, and (if at all) recreated in the future; it records opinions which may vary based on the work, the artist, the interviewer and date, about how to preserve variable media. MANS metadata is derived from the VMQ interview(s) which separates logical information from physical hardware. It is a metadata framework that accommodates description of physical and digital assets at the same level. Rinehart uses the metaphor of a musical score, as a form of declarative and conceptual notation of music, and likens Media Art to musical compositions that are able to maintain their original integrity while being realized by different instruments or in different arrangements, over evolving time periods. Ippolito's VMQ is a very complex method for focusing an artist's attention on those aspects of the work which need to be most enduring. The VMQ and MANs were designed to serve the needs of documentation, recreation and preservation of variable media.

A crosswalk of the MANS notation system or ontology to ten library, art, and preservation metadata standards (METS, PREMIS, EAD, FRBR, PBCore, etc.) verified the need for a domain specific schema for structural and administrative metadata. This poster demonstrates how the heart of the process, the VMQ, collects essential preservation information that MANS supports in a unique, infinitely extensible structure. MANS is derived as an interpretation of the MPEG-21 standard, primarily; the Digital Item Declaration Language (DIDL) a type of XML that allows for greater, more granular descriptions of multi-component digital objects. The intent is for MANS to act as a backbone to the artwork by being specifically suggestive though not overly prescriptive of how best to delineate and then, later, reinterpret an artwork, and thereby address long-term preservation strategies for variable media art such as storage, emulation, migration, reinterpretation.

References

Bekaert, K., Hochstenbach, P. and Van de Sempel, H. Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library. V.9, No.11, November 2003, D-Lib Magazine http://www.dlib.org/dlib/november03/bekaert/11bekart.html

Forging the Future Alliance. http://forging-the-future.net/

Functional Requirements for Bibliographic Records, Final Report. IFLA Study Group on the Functional Requirements for Bibliographic Records. http://www.ifla.org/VII/s13/frbr/frbr.pdf

Introduction to Metadata Pathways to Digital Information, Metadata Standards Crosswalks. Version 2.1, 2008. http://www.getty.edu/research/conduction_research/standards/intrometadata/crosswalks.pdf

Metadata Encoding and Transmission Standard: Primer and Reference Manual, Version 1.6 September 2007. http://www.loc.gov/standards/mets/METS%20Documentation%20final%20070930%msw.pdf

MPEG-21 Digital Item Declaration WD V2.0 http://xml.coverpages.org/MPEG21-WG-11-N3971-200103.pdf

PREMIS Data Dictionary for Preservation Metadata, Version 2.0. http://www.loc.gov/standards/premis/index.html

Ippolito, Jon. Death by Wall Label. http://vectors.usc.edu/thoughtmesh/publish/11.php

Rinehart, Richard. 20XX. A System of Formal Notation for Scoring Works of Digital and Variable Media Art. University of California, Berkeley.

User Guide to PBCore, Public Broadcasting Metadata Dictionary. Version 1.1. http://www.pbcore.org/PBCore/UserGuide.html.

Variable Media Network. http://www.variablemedia.net/

Geography of Impertinence: Using Historical Maps to Explore a Spanish Treatise on Piracy

Clayton McCarl
The Graduate Center of the City University of New York
clayton.mccarl@gmail.com

In recent volumes of *Literary and Linguistic Computing,* Martyn Jessop has considered digital visualization as a scholarly practice (2008b), and argued, in particular, for the expanded use of geographical information in digital humanities scholarship (2008a). These approaches hold much promise for literary and cultural studies, where attention has recently been given to such topics as "prose cartographies" (Padrón, 2004), the "cartographic imagination" (Smith, 2008), and a broad spectrum of other intersections between cultural production and physical space (both real and imagined), and especially within contexts of migration, conquest and colonialism (Arias, 2002; Brückner, 2007; Michelet, 2006). With *A Geography of Impertinence*, I hope to contribute to these discussions, raising questions about the visualization of geographical material in a seventeenth-century Spanish text, relationships between textual data and historical maps, and the potential for employing place names as a mechanism to facilitate cross-textual readings.

This project has grown out of my doctoral dissertation (currently in progress), an edition of *Piratas y contrabandistas de ambas Indias* (Pirates and Smugglers of the East and West Indies) by Spanish navigator and privateer Francisco de Seyxas y Lovera (1650-c.1705). *A Geography of Impertinence* responds to three central problems I face in the editing of this text.

The second question driving *A Geography of Impertinence* is the relationship between Seyxas' text and a 1630 atlas by the Portuguese cartographer João Teixeira Albernas. In *Piratas y contrabandistas*, Seyxas discusses this book, which he possessed for several years before presenting it to Charles II and the Council of Indies. The maps themselves contain Seyxas' handwritten additions and annotations, and include a rendering by Seyxas himself, disputing Teixeira's portrayal of Tierra del Fuego and the Straits of Magellan. Seyxas even inscribes himself in this geography, designating his own chain of islands: the "Islas de Seyxas." While these images could be used as illustrations, their close relationship to the text

– a relationship that is, in a sense, reciprocal – suggest that a "one-way" navigation from text to map is unnecessarily limiting. Jessop describes a visualization as being a "parallel (rather than subordinate) rhetorical device" (2008b, 283); certainly Teixeira's maps can serve this function in relation to *Piratas y contrabandistas*. Indeed, the interaction between map and text is a compelling question on its own, apart from all geographical considerations.

Lastly, I seek to understand Seyxas' book within a larger corpus of printed works dealing with issues of exploration and imperial competition. Just as *Piratas y contrabandistas* is tied to Teixeira's maps, it contains references to over 70 Spanish, English, Dutch and French books, mostly historiographical treatises and expeditionary journals. These include documents produced in the travels of individuals like Jacob Le Maire, Joris van Speilbergen, Bartholomew Sharp and John Narborough, as well as two volumes on nautical matters by Seyxas himself, printed in 1688 and 1690. These represent, of course, only a fraction of the enormous output of such books in the sixteenth and seventeenth centuries. While the intertexual bonds between these works go far beyond a common geography, place names provide a useful mechanism for exploring their differing geographical visions. For instance, the ability to navigate between these writings on the basis of toponyms may help us evaluate the role national or cultural bias plays in the imagination of place.

My presentation will demonstrate a prototype application which seeks to address the concerns outlined here. Built around a collection of place names, this application has much in common with a digital gazetteer, and thus draws upon work done by such projects, including the *Alexandria Digital Library Gazetteer*. The instances of a given place can be either textual or cartographic, and exploration of the collection can begin either at the index or within the texts or maps themselves. The model allows for the inclusion of a scholarly apparatus, as recommended by Jessop (2008b, 291) and as encouraged by the principles for documentation discussed in the *London Charter for the Computer-Based Visualization of Cultural Heritage* (8-9). This apparatus is implemented using an object-oriented approach, enabling the sharing of editorial material across textual and cartographic objects, and facilitating different levels of granularity in annotation, depending on the context of a given place instance. The model permits multiple names for a given place, and allows for annotation in multiple languages. I offer as test cases some of the most problematic places mentioned by Seyxas, and include in my textual corpus a sampling of passages by authors cited in *Piratas y contrabandistas*.

References

Alexandria Digital Library Project. (1999-). Gazetteer Development. Available from: http://www.alexandria.ucsb.edu/gazetteer (accessed 12 March 2009).

Arias, S. & Meléndez, M., eds. (2002). *Mapping Colonial Spanish America: Places and Commonplaces of Identity, Culture, and Experience*, Lewisburg, Pa.: Bucknell University Press.

Brückner, M and Hsu, H., eds. (2007). *American Literary Geographies: Spatial Practice and Cultural Production, 1500-1900*, Newark: University of Delaware Press.

Jessop, M. (2008a). The Inhibition of Geographical Information in Digital Humanities Scholarship. *Literary and Linguistic Computing*, 23(1), 39-50.

Jessop, M. (2008b). Digital Visualization as a Scholarly Activity. *Literary and Linguistic Computing*, 23(3), 281-293.

The London Charter (2009). London Charter for the Computer-Based Visualization of Cultural Heritage (Draft 2.1). Available from: http://www.londoncharter.org (accessed 12 March 2009).

Michelet, F. (2006). *Creation, Migration, and Conquest: Imaginary Geography and Sense of Space in Old English Literature*, Oxford: Oxford University Press.

Padrón, R. (2004). *The Spacious Word: Cartography, Literature, and Empire in Early Modern Spain*, Chicago: University of Chicago Press.

Smith, D.K. (2008). *The Cartographic Imagination in Early Modern England: Re-writing the World in Marlowe, Spenser, Raleigh and Marvell*, Aldershot, UK: Ashgate.

Capturing the Social Networks of the Gospels through a Graph Clustering

Maki Miyake
Osaka University
mmiyake@lang.osaka-u.ac.jp

The creation of social network representation is a promising application of large-scale linguistic resources as a means of capturing the patterns of social relationships that exist among the individuals. In recent years, several notable studies have produced a number of social networks of the Bible and a variety of its graph representations for gaining insights into the interactions between the characters (Chris Harrison 2007, ESV Blog, 2007). Graph representation is an effective way of detecting and investigating the intricate patterns of connectively within large-scale corpora. A number of studies have applied graph theory and network analysis methods to mapping out the complex networks of word associations within linguistic resources (Dorow, Widdows, Ling, Eckmann, Danilo, and Moses, 2005; Steyvers and Tanenbaum 2005; Gfeller, Chappelier, and De Los Rios, 2005). In terms of the biblical texts, I have successfully applied a graph clustering technique to data processing that utilizes a clustering-coefficient threshold in creating sub-networks for the Gospels of the New Testament (Miyake, 2008). This paper reports on the application of a soft graph clustering method to four social networks of the Gospels constructed based on the co-occurrences of the people and places. Specifically, I propose a soft graph clustering technique as a method of detecting the community structures within the social networks of the Gospels. The principle objectives of this study are to investigate the interaction between the characters in the stories.

The corpus used in the present study is the Greek version of the Gospels (Nestle-Aland, 1979) and the data mainly consists of names and places. A set of nouns such as son, father, prophet are including as well, that are regarded as important in representing the characteristics and the roles of people in the stories. In creating four social networks from the books of Mark, Matthew, Luke and John, co-occurrence data is computed for pairs of words that appear in the same verse (sentence) and the words remain morphological forms that make is possible to analyze the relationships between words such as "who says whom".

Table 1 presents the number of co-occurrence words represented as nodes for each Gospel network and its basic statistical measure such as degree average values, the average shortest path and the average of clustering coefficient that are clues for examining the structural characteristics. Degree refers to the number of words that are connected to a given word and its average value shows its connectedness of nodes within a network. These degree average values for each Gospel indicate that these social networks have patterns of sparse connectivity. The clustering coefficient is known as the index for investigating probabilities that an acquaintance of an acquaintance is also an acquaintance of yours, in other words, as an index of the inter-connective strength between neighboring nodes in a graph. Following Watts and Strogatz (1998), the clustering coefficient of the node (n) can be defined as: (number of links among n's neighbors)/(N(n)*(N)-1)/2), where N(n) denotes the set of n's neighbors. The coefficient assumes values between 0 and 1. In the social network study, the clustering coefficient can be utilized as a measure of role ambiguity or community hubs that have numerous links. The four average clustering coefficients for the total nodes are all quite high values of 0.5 or more, indicating strong connectedness between nodes, which is a characteristic of small-world networks.

	Mk	Mt	Lk	Joh
node	258	179	257	146
Degree average	5.09	5.92	5.3	6.96
(% for the total nodes)	2%	3%	2%	5%
Average<C>	0.58	0.55	0.57	0.52

Table 1

In order to discern the relationships among the words within the social network, this study applies a soft clustering method combining the hard clustering of Markov Clustering (MCL) and the index of clustering coefficients. MCL is the bottom-up classification methods that allow us to detect the patterns and clusters within large and sparsely connected data structures. Within the MCL process, a graph is partitioned into hard clusters. However, one particular problem with applying MCL to linguistic resources is that the hard clustering approach is not appropriate for words that have multiple meanings, more specifically for the individuals who are involved in multiple communities in the context of social networks. In order to overcome this problem, that hard clustering of

MCL is applied to the network removed the bottleneck nodes which are identified as the hubs by taking the clustering coefficient as a threshold. In this study, the only nodes with the clustering coefficient of more than 0.2 were selected. After the MCL process, the resultant crisp clusters are expanded with neighboring nodes to produce overlapping clusters that include in the hub nodes.

Figure 1 plots the numbers of the MCL cluster sizes for each network, which illustrate the transitions occurring in downsizing the networks generated from graph clustering. Taking the result of Mt for an example, the MCL resulted in 44 hard clusters, with an average of ?? cluster components (SD=??). Comparing between clusters across the four social networks illustrates the different roles associated with characters in the Gospels, such as Jesus as the son of God in Mark and Jesus as Savior in Luke.

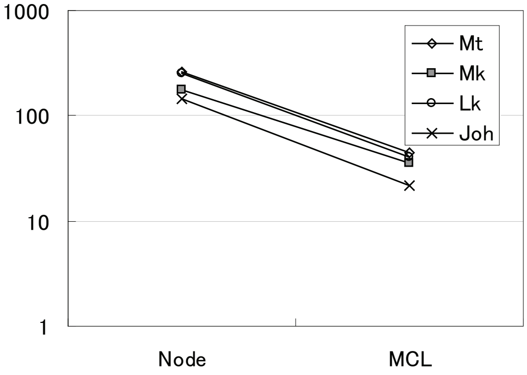


Figure 1

In order to compare between clusters across the four social networks, the Jaccard coefficient is appropriate index for measuring similarity between clusters, which can be computed based on the number of elements in the intersection set divided by the number of elements in the union set. The table 2 presents the results of average Jaccard coefficient value among four Gospels, indicating that the similarities among first three Gospels such as Matthew, Mark, and Luke are higher than those for John. The first three Gospels have been referred to as the Synoptic Gospels, because a high similarity has been recognized among them. Table 3 presents a set of clusters with the highest similarity for the Synoptic Gospels, which can easily refer to a phrase beginning with "Render onto Ceasar". Figure 2 is a sample of the sub-networks focusing on the word of Peter and its neighboring nodes. As the neighboring structures for the node of Peter are different among four networks, such a graph representation makes it possible to examine structural similarities and differences among social networks.

	Mt	Mk	Lk	Joh
Mt	1	**0.198**	**0.215**	0.128
Mk	–	1	**0.195**	0.125
Lk	–	–	1	0.109
Joh	–	–	–	1

Mt	Mk	Lk
{θεου(God:gen), θεω(God:dat), καισαϱι(Ceasar: dat), καισαϱος (Ceaser: gen)}	{ανθϱωποις, θεου (God:gen), θεω (God:dat), ιησους (Jesus:nom), καισαϱι(Ceasar: dat), καισαϱος (Ceaser: gen) }	{ανθϱωποις, θεω (God:dat), καισαϱι (Ceaser: gen)
22:21 They say unto him, Caesar's. Then saith he unto them, Render therefore unto Caesar the things which are Caesar's; and unto God the things that are God's	12:17 And Jesus answering said unto them, Render to Caesar the things that are Caesar's, and to God the things that are God's.	20:25 And he said unto them, Render therefore **unto Caesar** the things which be **Caesar's**, and **unto God** the things which be **God's**.

Table 2

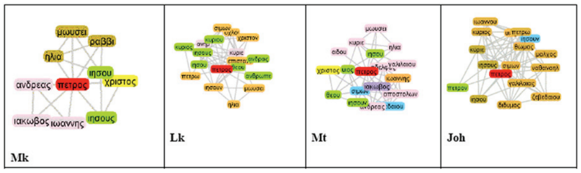


Figure 2

In summary, this paper has reported on the application of a soft clustering method combining the clustering coefficient and Markov Clustering. Especially, the graph clustering technique offered an effective way of controlling over the hub nodes that are linked to numerous other nodes. Examining social networks is useful in exploring the interactions between characters and the features that underlie word groups within the Gospels. In pursuing the precise communities the characters are involved in and the series of action and events in stories, the research is working to make the dataset more sophisticated, such as the treatment ambiguous names and personal pronouns.

References

S. van Dongen. (2000). *Graph Clustering by Flow Simulation*, PhD thesis, University of Utrecht.

B. Dorow, D. Widdows, K. Ling, J. Eckmann, D. Sergi, and E. Moses. (2005). Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination, *Proceeding of 2nd Workshop organized by MEANING Project (MEANING-2005)*.

M. Steyvers, and J. B. Tenenbaum. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth, *Cognitive Science*, 29 (1): pp.41-78.

Gfeller, D., Chappelier, J.-C., and De Los Rios, P.. (2005). Synonym Dictionary Improvement through Markov Clustering and Clustering Stability, *International Symposium on Applied Stochastic Models and Data Analysis*, pp. 106-113.

Maki Miyake (2008). Investigating word co-occurrence selection with extracted sub-networks of the Gospels Employing Clustering Coefficients, *Digital Humanities 2008*, pp.258-260.

ESV Blog. (2007). Mapping New Testament Social Networks, <http://www.esv.org/blog/2007/01/mapping.nt.social.networks>

Chris Harrison. (2007). Visualizing the Bible, <http://www.chrisharrison.net/projects/bibleviz/index.html>

Aspects of the Interoperability in the Digital Humanities

A Case Study in Buddhist Studies

Kiyonori Nagasaki
Yamaguchi Prefectural University, Japan
nagasaki@ypu.jp

A. Charles Muller
The University of Tokyo, Japan

Masahiro Shimoda
The University of Tokyo, Japan

1. Introduction

In considering cases of interoperability in the Digital Humanities, it might be useful to focus on an example from Buddhist Studies for three reasons: The first is usage of the multilingual Buddhist canons that were written in Pali, Sanskrit and translated into Chinese, Tibetan, and so on before the 10th century. Thus, a system had to be established that allowed one to deal with resources that are composed not only by various families of written characters but also various languages. The second is that they are significant as resources in history, linguistics, and other academic fields because they include not only Buddhist ideas but also information related to the ancient world of which Buddhism was a part. Thus, interoperability in this field is needed in order to synthesize such related fields. The third reason to focus on this case study is that digitization projects in the field Buddhist Studies are in progress worldwide, having both academic and proselytizing purposes. Moreover, many projects house sub-projects within: for example, one's aim might be to merely retrieve information on a local computer; or, on the other hand, one might want to publish critical editions of digitized canonical texts on the Web. Discussion of this need for interoperability has already begun in the organization known as the Electronic Buddhist Text Initiative [EBTI]. We will discuss some aspects of that through our case study that follows.

2. A Case Study: SAT and the DDB

In this chapter, we will discuss the interoperability between two different projects in which we are engaged: the DDB (Digital Dictionary of Buddhism, *http://www.buddhism-dict.net/ddb/*) and the SAT (SAT Daizōkyō Text Database Committee, *http://21dzk.l.u-tokyo.ac.jp/SAT/*). The DDB is a web-serviced lexicon that includes over 45,000 entries. The SAT project has digitized a scholarly edition of the Chinese Buddhist canon, consisting of approximately 150 million Chinese characters in eighty-five volumes. It was compiled and edited by Japanese scholars in the Taishō Era and has been treated as de facto standard text in the field of Buddhist study since then.

2.1. The DDB: The Digital Dictionary of Buddhism

The DDB was developed for the study of Buddhist texts written in classical Chinese and other East Asian languages that include Chinese character-based terminology. This project was initiated in 1986 by Charles Muller, a specialist in East Asian Buddhism. In 1995, with the advent of the Internet, Muller converted his data set into HTML format, and placed it on the Web. During this period from 1996-2000, the storage format of the dictionaries was changed to SGML, and then to XML. In 2001, with the help of humanities computing guru Michael Beddow, the dictionaries were reset on the web in XML format with a search engine, and this structure remains in place down to the present day. The DDB features Buddhist terms, texts, schools, temples, and persons. Entries range in scope from short glossary type, to full-length encyclopedic articles. Now supported by more than sixty collaborators with specialist's expertise in a wide range of areas in Buddhist studies, the expansion rate of the DDB has been exponential in recent years.

A special dimension of the DDB is its usage of XML attributes to accredit contributors for their work at the level of entry sub-areas (XML "nodes") rather than only at the level of full entries, as seen in standard printed works. Furthermore, the relatively accessible XML tag structure (based loosely on the TEI model) has made it possible to integrate and interlink the DDB with other lexicons (such as EDict), and external text databases, such as the SAT text database.

2.2. SAT: The SAT Daizōkyō Text Database Committee

SAT is managed by the SAT Daizōkyō Text Database Committee (directed at present by Masahiro Shimoda). The database depends on an XML-like legacy scheme which was designed in 1998 and superficially represents the pages of the edition. The textual corpus was digitized and corrected mainly by about 200 young Buddhist scholars during a period of about ten years. While descriptive markup has not yet been fully implemented, the locations of passages in the source texts, such as page, paragraph and line were precisely recorded so that the traditional methodology of the Buddhist study could be referred to transparently. It has been posted on the Web since April 2008.

2.3. The Interoperation between the DDB and the SAT

When the SAT Web service was started, it provided some interoperation with other related projects. The most important service provided is the search function for the entries of the DDB. The function adopts AJAX so that users can retrieve items transparently on their Web browsers. If users select a portion of the text with their mouse devices, they can view all terms in the text contained in the DDB. The service is convenient for those who have interest in the text—especially beginners in Buddhist studies. In addition, the SAT Web service distributes some APIs. One of them is reference service based on the physical location. It provides a function to clip an arbitrary part of the texts by means of specifying the location or the range in a certain URI. The DDB and some other Web services adopt this API. The important merit of this interoperation is not only the usability, but also the management structure that allows each long-term project to be sustained independently. It is so difficult to manage a big digital project that sometimes it may be disrupted, crash or even disappear. Although we may wish to increase our services, they often become unwieldy, even ending up in abandonment of the project. Enhancing Web services in order to support researchers, the interoperation with the other projects will also be a workable alternative in the field of Buddhist studies.

3. Other Examples of Interoperation

SAT interoperates with some other projects such as CHISE and INBUDS. CHISE is an ontology mainly focusing on Chinese characters distributed under GNU GPL. SAT adopted CHISE to serve as a thesaurus of Chinese characters in order to support its retrieval system.

INBUDS is a bibliographical database for the study of Indian philosophy and Buddhism. It is maintained by the Japanese Association of Indian and Buddhist Studies and includes 60,000 records that have been collected for twenty years. SAT implemented the interoperation with the INBUDS so that users could refer to the related academic resources. Some of the resources in the INBUDS are distributed as digital data.

4. Some Problems in Interoperability

As discussed previously, in the field of the Buddhist studies, interoperability is quite efficient in some ways. On the other hand, all-too-common problems of interoperability are also found in the field. One of the important issues is that of organizational sustainability. If one side of the organization stops distribution of their own resources, the interoperation would end. However, it can be argued from another perspective that this is actually an advantage, because it allows ready awareness of systematic changes, allowing managers to adapt as necessary, for example, when one needs to salvage distributed data. Indeed, SAT has recently begun to support INBUDS because, after establishing the interoperation, it became clear that the programmers of INBUDS had been facing some problems with managing their data for some time. Therein, especially, in the case of a personal project, it is more secure to establish interoperation.

5. Conclusion

Just as “the fourth generation collections” in the field of the Western classics, digitizing projects for Buddhist studies are gradually shifting their own styles to the next generation which puts emphasis on interoperability. Interoperability not only exposes the problems inherent in the activities of the digitization of Buddhist studies, but also shows the ways to solve those problems. The same model will eventually hold true for the Humanities in general.

References

Muller, A. Charles. (2008). EBTI After 15 and CBETA after 10 Years: Joint International Conference on Digital Buddhist Studies, Chair's Report.

http://buddhism-dict.net/ebti/ebti2008report.html (accessed 12 November 2008).

Muller, A. Charles. (2008). The Digital Dictionary of Buddhism [DDB]: Present Status and Future Developments, The Ninth Annual Symposium for Scholars Resident in Japan, March 2008. *http://www.acmuller.net/articles/ddb-nichibunken-200803.html* (accessed 12 November 2008).

Crane, G. (2008). Fourth Generation Collections: TEI, FRBR, and Canonical Text Services, TEI Member’s Meeting 2008, Nov 2008. *http://www.cch.kcl.ac.uk/cocoon/tei2008/programme/abstracts/abstract-160.html* (accessed 12 November 2008).

Rehm, G. and Witt, A. (2008). Aspects of Sustainability in Digital Humanities, Digital Humanities 2008 , June 2008: 21-29.

Nagasaki, K. and Shimoda, M. (2008). Outline of the Activities of the SAT Project, *Joint International Conference on Digital Buddhist Studies, at Dharma Drum Buddhist College*, February 2008: 22-23.

Nagasaki, K. (2008). A Collaboration System for the Philology of the Buddhist Study”, *Digital Humanities 2008*: 262-263.

MORIOKA, T. (2006). Character processing based on character ontology, *IPSJ Technical Report,* 2006-CH-072: 25-32.

Eide, Ø., Ore, C. and Holmen, J. (2008). Sustainability in Cultural Heritage Management, *Digital Humanities 2008* : 22-23.

The *Soweto ’76* Archive: Virtual Heritage, Human Rights & Social Justice in the New South Africa

Angel Nieves
Hamilton College
anieves@umd.edu

The Maryland Institute for Technology in the Humanities (MITH) and the Hector Pieterson Memorial & Museum (HPMM) propose a transatlantic digital collaboration to create the Cultural Heritage Platform, an extensible web interface and toolset for the detailed study and conservation of historic resources. We further propose an extensive technical training and support program for curators and staff at Constitution Hill, Johannesburg; The District Six Museum, Cape Town; Red Location, Port Elizabeth; and Kliptown, Johannesburg, enabling them to populate the digital archive with content from their collections. This joint effort would create a core “digital cultural heritage trail” that could be further extended to other related South African museums and that would allow students, teachers, and the general public from anywhere in the world to explore digital recreations of some of the most important places in the struggle against Apartheid. This paper will focus on the development of a digital archive, *Soweto ’76*, currently being built at the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland in collaboration with the Hector Pieterson Memorial & Museum.

The *Soweto ’76* project began in early 2006 as a collaboration between the Maryland Institute for Technology in the Humanities (MITH) and the Hector Pieterson Memorial & Museum (HPMM), Johannesburg, South Africa. The initial scope of the project included the digitization and preservation of the archival collections of the Museum with the intention of providing on-line access to its holdings for broad public use. Due to a lack of available resources for their care and preservation, these holdings were considered endangered, and the MITH project team began the process of digitizing them in 2006. By early 2007 the team determined that the material in the archive could be best presented in an interactive 3D virtual environment with social networking functionality. Such an environment would stimulate critical historical thinking by raising questions about the nature and construction of historical narratives in newly developing democracies. In addition, the online archive would make the materials in it broadly available to scholars interested in doing research on the history of Soweto and the student movement against Apartheid. Currently *Soweto ’76* is completing the digitization of its holdings including oral history interviews, video interviews, historic photographs, commemorative memorabilia, maps, and other material artifacts. A collection of over 200 South African newspaper articles from the period 1976-1980 have also been digitized and transcribed as part of the archive. The project staff has recently started the process of “tagging” (that is, describing) text, image, video, and audio files from already digitized collections.

As MITH developed the *Soweto ’76* interface, they worked, according to the best principles of software design, to make their work as generalizable and as reusable as possible. It soon became apparent that the interface might be used for other archives to create a larger cultural heritage platform for historic sites. This large, multi-institutional archive has become the goal of the project we are now proposing.

The link between human rights and the preservation of cultural heritage resources is often misunderstood. Forgetting our histories is politically problematic for many reasons, primarily because it denies the potential for building broader cultures of democracy. Nations must apologize and/or offer compensation for historical injustices if there is to be a process of healing and remembrance. If we are truly seeking social justice, we must remember these historical injustices and recognize how they continue to shape identities even today. It is therefore essential to understand cultural heritage resources as a part of peoples’ efforts to maintain and construct their own identity. Historic sites are critical elements in the struggle for equality and democracy, and new technologies can be used to increase access to the information kept in these important spaces.

In addition to providing access to archival materials extant in collections such as those of HPMM, our database interface will allow users to explore historically accurate recreations of heritage spaces in South Africa, and to access digital artifacts related to each particular space/place. This focus on place is important because so much of the critical urban fabric of places such as Soweto, District Six, or Red Location was erased during the Apartheid era. Under the post-apartheid Mandela and Mbeki governments there has been an assumed connection between urban redevelopment and heritage programs, particularly if they promote foreign tourism. As some scholars have argued, “political transformations can create new urban political identities, but the politics of tourism can give cash value to the memorializing of select pasts." Over a decade of democracy has brought massive reforms and advances across the heritage industry, but there remains a lack of understanding regarding the cultural significance of Black heritage resources in South Africa's still isolated townships. Unfortunately, much of the physically extant heritage does not easily lend itself to the traditional standards of what is considered "architecturally significant" or "visually impressive." Much of the history of the anti-apartheid movement took place in the townships amongst what many heritage professionals would consider to be "the mundane" and ordinary structures and environments of the poor. Therefore, it is necessary to develop a different set of criteria and strategies for documenting and preserving these important sites.

Once the Soweto interface is complete, other cultural heritage institutions in South Africa concerned with human rights and social justice can be added easily to the digital heritage trail, as long as they conform to our metadata standards. Each additional institution thus will not only make its own content visible to the world, but will also enrich the collectively searchable content of the archive as a whole, which will always be greater than the sum of its constituent parts. In this way we hope to create a "Digital Heritage Trail" that connects the cultural institutions in South Africa together in virtual space.

Of course, creating the XML files that comprise each digital object can be demanding work, especially for those whose technical skills are not highly developed. For this reason, we will also provide access to the Ajax XML Encoder (AXE), developed in 2007-2008 at MITH (originally for the *Soweto '76* project) with funding from a National Endowment for the Humanities Digital Start-up grant. This tool allows non-technical users to describe or "tag" text, image, video, and audio files using an intuitive Web-based interface. Graduate students working on the Soweto project have already successfully used the tool, and a public beta version is planned for the end of the summer of 2008. Because our interface will be open source and our metadata standards will be readily available, they will be adaptable for the collections of other cultural heritage institutions in South Africa and around the world.

Institutional Partners

Hector Pieterson Museum & Memorial (HPMM), Johannesburg, South Africa Maryland Institute for Technology in the Humanities (MITH), College Park, MD, USA Africana Studies Program, Hamilton College, Clinton, New York, USA

Project Partners

Dr. Angel David Nieves, *Soweto '76* Project Director
Mr. Ali Khangela Hlongwane, Chief Curator & Museum Director HPMM
Dr. Doug Reside, MITH Assistant Director
Mr. Greg P. Lord, MITH Web Designer & Software Engineer
Mr. Arik Lubkin, Graduate Student, UM-CP

Close Only Counts in Horseshoes and... Authorship Attribution?

John Noecker Jr.
Duquesne University
jnoecker@gmail.com

Mike Ryan
Duquesne University
michaelryan@acm.org

Patrick Juola
Duquesne University
juola@mathcs.duq.edu

Amanda Sgroi
Duquesne University
sgroia@duq.edu

Stacey Levine
Duquesne University
sel@mathcs.duq.edu

Benjamin Wells
Duquesne University
wellsb1930@duq.edu

How much of an effect do transcription errors in a text document have on the ability to do useful statistical analysis on that document? In order to perform authorship attribution, it is often necessary to first have a digital copy of the documents in question. The task of authorship attribution is to assign an authorship tag to a document of unknown origin based on statistical analysis of the text and comparison with documents of known authorship. This is often automated by means of a computer, which necessitates the existence of digital copies of all the works to be analyzed. The success rates of optical character recognition (OCR) systems make them an attractive choice for the creation of these digital copies. The various documents can be scanned into a computer and converted automatically to text. Rice et al. documented OCR per-character accuracy rates of greater than 90% for nearly all commercial OCR systems tested, most of which scored in a range of 95-98% accuracy (1996). More recent commercial claims by OCR companies suggest that accuracy rates are above 98%. However, is a 5% or even 2% error rate acceptable when creating a statistical authorship model? Is it necessary to proofread each scanned document by hand before performing authorship attribution, or is the error rate small enough that it is unlikely to affect the overall result?

We intend to present new results showing that it is not necessary to proofread scanned documents before using them to perform statistical authorship attribution. In fact, these results suggest that no significant performance degradation occurs when analyzing documents with per-character error rates of less than 15%. As this is well below the published averages for OCR systems, there is little need to worry about the few errors which will be introduced during the automated image-to-text conversion. Accuracy of study materials is of course crucial for an ideal authorship attribution study. As Rudman (1997) put it, “most non-traditional authorship attribution researchers do not understand what constitutes a valid study.” In 2002, he wrote, “do not include any dubitanda —a certain and stylistically pure Defoe sample must be established—all decisions must err on the side of exclusion. If there can be no certain Defoe touchstone, there can be no ... authorship attribution studies on his canon, and no wide ranging stylistic studies.” Ideally, we would have access to the original manuscripts to make sure that what we have is the pure work --but in the real world, we may not have such access. We argue, in contradiction to Rudman, that by assessing the likely contribution of types of error—such as errors introduced by bad OCR technology—we can determine whether such errors are likely to shake our confidence in our results. If we can give 10:1 odds that a given paper was likely written by Defoe, we will still be confident if we learn that our odds might be as low as 9.8 or 9.9:1.

For this experiment, we made use of the Java Graphical Authorship Attribution Program (JGAAP *www.jgaap.com*), a freely available Java program for performing authorship attribution created by Patrick Juola of Duquesne University. This modular program breaks the task of authorship attribution into three subtasks, described as ’Canonicization’, ’Event Generation’ and ’Statistical Analysis’. During the Canonicization step, documents are standardized and various preprocessing steps can occur. For this experiment, we created a Canonicization method which randomly changed a percentage of the characters in each document, simulating the per-character error of an OCR system. We also converted all characters to lower case during this step. We generated a feature set of ’Words’, which JGAAP defines as any string of characters separated by whitespace. Finally, we used a normalized dot product as a nearest neighbor algorithm to assign authorship tags to unknown documents. Noecker and Juola (personal correspondence) have suggested that this normalized dot product scoring, which they refer to as the ’Cosine Distance’, outperforms many more complicated techniques and is especially well suited to the feature set we chose.

In order to test this experiment on real world data, we have used the Ad-hoc Authorship Attribution Competition (AAAC) corpus. The AAAC was an experiment in authorship attribution held as part of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities. The AAAC corpus provides texts from a wide variety of different genres, languages and document lengths, assuring that the results would be useful for a wide variety of applications. The AAAC corpus consists of 98 unknown documents, distributed across 13 different problems (labeled A-M). An analysis method’s AAAC score is calculated as the sum of the percent accuracy for each problem. Hence, a AAAC score of 1300% represents 100% accuracy on all problems. This score was designed to weight both small problems (those with only one or two unknown documents) and large problems equally. Because this score is not always sufficiently descriptive on its own, we have also included an overall accuracy rate in our experiment. That is, we calculate both the AAAC scoring and the total percentage of unknown documents which were assigned the correct authorship labels. These two scores provide a fair assessment of how the technique performed both on a per-problem and per-document basis.

The experiment itself consisted of fifty-two iterations of the authorship attribution process over the entire AAAC corpus. The presented results will be an average of the effect of random transcription errors from 1% to 100% of the characters in each document. As previously noted, we calculated both the AAAC score and overall percentage correct over some 5,096 experiments (98 experiments per iteration). These overall results suggested that there was essentially no decrease in performance on either the AAAC score or the overall percentage for error rates of 1% to 2%, which many commercial OCR systems claim to achieve. Even for more skeptical error rates of 5%, the overall percentage correct decreased by about 1%, and the AAAC score by only about 20. This trend continues until roughly a 15% error rate, after which the performance drops off rather considerably. Still, as Rice et al. have reported, all major OCR systems are capable of error rates considerably lower than 15%, which strongly suggests that there is little reason to spend additional time proofreading scanned documents before performing authorship attribution.

Graceful Degradation: Managing Digital Projects in Times of Transition and Decline

Bethany Nowviskie
University of Virginia
bethany@virginia.edu

Dot Porter
Digital Humanities Observatory
dot.porter@gmail.com

Past panel discussions at Digital Humanities conferences (including one on the question of “finished” work and another on innovative management techniques) have offered anecdotal evidence about factors contributing to the success of digital projects.[1] Our journals and conference programs brim with accounts of work in progress, generally presented at the height of its success. The exigencies of grant funding and conference or publication submissions tend to make the record more sanguine about our projects than, perhaps, their full lifecycle would merit. This poster takes a deliberative look at that darker side of project management with which we are all too familiar – the experience of projects that have entered states of transition or even decline. We will open and describe a research methodology for a broad survey of the digital humanities community related to project management – specifically on how we think about our projects and behave toward them when they face times of transition and decline, and what we see as the causes and outcomes of those times.

Decline is an especially pressing issue for the digital humanities because of the tendency of our projects to be open ended. Traditional, long-term scholarly projects (even the small subset which, like the majority of digital projects, are collaborative in some way) are generally projected to end with the publication of a monograph or scholarly edition – something solid to sit on the shelves, regardless of whether the project itself endures past a publication point. Digital humanities projects are more likely to have output such as databases or websites, objects that beg a sustainability plan and require long-term curation even if they are not continually updated.[2] They are also more collaborative than most scholarly projects and therefore more dependent on the continuing energy and goodwill of various institutional and intellectual partners. One could argue that digital projects are, by nature, in a continual state of transition or decline.[3] What happens with the funding runs out, and the original project staff has moved on or been replaced? What happens when intellectual property rests with a collaborator or an institution that does not wish to continue the work?

If projects are particularly vulnerable in transitional phases, how can we anticipate and ameliorate the effects of these times? Does the valuation of projects against conventional measures of scholarly success (teaching, research, service) or within traditional disciplinary boundaries impact their continuity or conclusion? Is there actually something qualitatively different about digital projects versus "traditional" scholarly undertakings? Are there new models for scholarly output that match more satisfactorily to the real-world outcomes and trajectories of digital projects? Do certain kinds of early planning make projects more likely to weather changes? What brands of institutional support are most helpful to projects that are meeting their natural or unnatural ends?

Survey questions address the following issues:

- basic questions of project funding (its duration and dependability) and the role of local, institutional support;
- the early definition of the project and its potential for "mission creep;"
- the definition of short-term vs. long-term goals for the project and to what extent the project met them;
- the relation of the project to inquiry in traditional disciplines, to pedagogy, to published research, and to tenure and professional advancement;
- staff retainment and continuity;
- matters of intellectual property and open source;
- the use of tools or techniques for project management;
- explicit risk management in project design and the "sustainability" both of the product of the work and of its production process;
- and whether the views of survey participants regarding digital scholarship have evolved over time in response to transitional experiences with their own projects or to larger changes in academic culture.

The "Graceful Degradation" survey has been constructed under advisement of experts in qualitative data analysis at the University of Virginia. It will be conducted online and in paper format over the course of several months beginning in the summer of 2009. The authors of the survey are well positioned to solicit responses from both North American and European projects and will endeavor to reach out more broadly – both geographically and to communities of practice (including digital history, computer music, and electronic publishing) that have been somewhat underrepresented at past digital humanities gatherings.

At Digital Humanities '09 in Maryland, we will solicit responses from conference attendees and present our methodology in poster format. We anticipate that this survey will help us to determine fruitful lines for future inquiry, including projects deserving of careful case study presentation. We hope to identify and share some best practices in the design and management of projects that weather transitional periods well. We would also deem this poster presentation a success if it broke the taboo on conversation about those digital humanities efforts that – for reasons we will be prepared to describe – have failed to degrade with a measure of grace.

Notes

[1] "Innovations in Interdisciplinary Research Project Management," Ramsay et.al., *Digital Humanities 2008* (Oulu, Finland) and "Done: 'Finished' Projects in the Digital Humanities," Kirschenbaum et.al., *Digital Humanities 2007* (Urbana-Champaign, Illinois).

[2] Daniel Pitti, "Designing Sustainable Projects and Publications," in *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, 2004. http://digitalhumanities.org/companion/ (accessed 13 November 2008).

[3] Diane Zorich, "A Survey of Digital Humanities Centers in the United States." CLIR Reports, November 2008. http://www.clir.org/pubs/reports/pub143/contents.html (accessed 13 November 2008).

Markup Schemes for Manga and Digital Reformatting Systems

Kazushi Ohya
Tsurumi University
ohyakazushi@gmail.com

1 Introduction

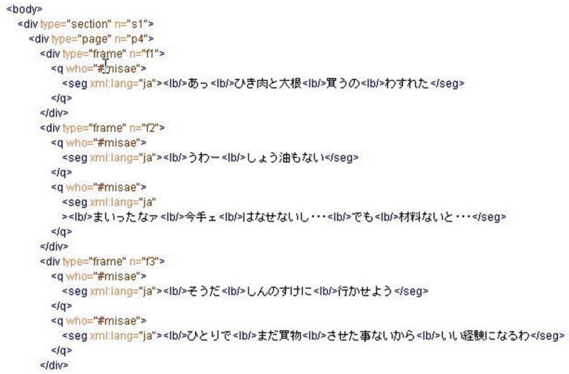
In this paper, we will present two schemes, one for encoding Manga and one for the collation of graphic resources, both of which are used in reformatting systems. The schemes and reformatting systems will provide multiple merits for carving out a new research field for digital humanities.

2 New Field Opened Up by Manga

Manga, or a graphic novel originating in Japan is gradually becoming known as new reading art, or a kind of literature even outside of Japan. Manga contains multiple graphic units on a surface unit(like a page) and sometimes texts are written in the background, which we call depicted letters. Each frame, which is a unit of an image, is sometimes related to another frame using a caption. And, each balloon can be dominated by multiple frames and possibly other balloons. It means that we have to prepare a scheme to handle graphic units with structures which are not the same as the text structure. To digitize Manga in markup languages requires more fine encoding schemes for the relationship between image and text data rather than those for simply illustrated texts, like referring directly a corresponding image file. Manga could be a new target for encoding in markup languages, and a good test bed for checking adaptability of reference schemes like TEI[1].

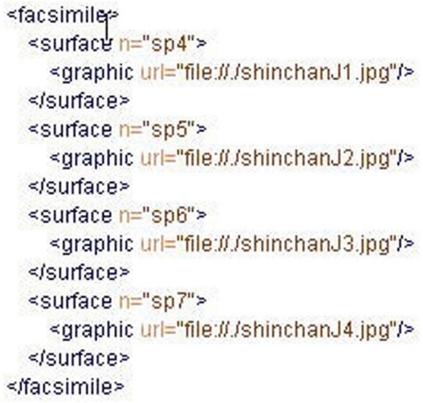
Manga has graphic expressions which help us envisage situations the story represents, which means that translations of Manga will not be so different from language to language, compared to translations of other types of literature like novels. Then, we can expect that the multi-lingual corpora based on Manga will be semantically more stable. The corpora would be good resources of analysis for natural language processing especially for machine translations.

Manga has the potential also for good corpora of onomatopoeia, or ways to imitate phenomena with letters, which can be regarded as a type of mimicry using phonetic values of letters/characters. Onomatopoeia have been neglected as academic research targets for a long time since its presence in linguistics or semiotics has been denied. However, in our view, onomatopoeia is a full-fledged domain of linguistics as well as nouns, sentences, and discourse units, or at least should be a domain for lexical items in multi-lingual dictionaries. Manga will provide a good resource for making corpora of onomatopoeia.



```
<body>
  <div type="section" n="s1">
    <div type="page" n="p4">
      <div type="frame" n="f1">
        <q who="#misae">
          <seg xml:lang="ja"><lb/>あっ<lb/>ひき肉と大根<lb/>買うの<lb/>わすれた</seg>
        </q>
      </div>
      <div type="frame" n="f2">
        <q who="#misae">
          <seg xml:lang="ja"><lb/>うわー<lb/>しょう油もない</seg>
        </q>
        <q who="#misae">
          <seg xml:lang="ja"
          ><lb/>まいったなァ<lb/>今手ェ<lb/>はなせないし・・・<lb/>でも<lb/>材料ないと・・・</seg>
        </q>
      </div>
      <div type="frame" n="f3">
        <q who="#misae">
          <seg xml:lang="ja"><lb/>そうだ<lb/>しんのすけに<lb/>行かせよう</seg>
        </q>
        <q who="#misae">
          <seg xml:lang="ja"><lb/>ひとりで<lb/>まだ買物<lb/>させた事ないから<lb/>いい経験になるわ</seg>
        </q>
      </div>
```

List 1



```
<facsimile>
  <surface n="sp4">
    <graphic url="file:///shinchanJ1.jpg"/>
  </surface>
  <surface n="sp5">
    <graphic url="file:///shinchanJ2.jpg"/>
  </surface>
  <surface n="sp6">
    <graphic url="file:///shinchanJ3.jpg"/>
  </surface>
  <surface n="sp7">
    <graphic url="file:///shinchanJ4.jpg"/>
  </surface>
</facsimile>
```

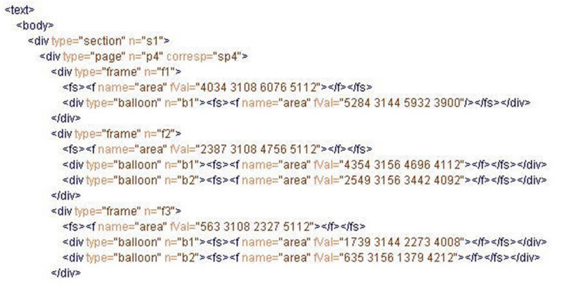
List 2

3 Schemes for Text and Images of Manga

As a preliminary experiment, we had encoded two titles of Manga, Nodame Cantabile and Crayon Shinchan, following the scheme of a drama module for performance text in TEI P5, and confirmed that the scheme can also be used as a base scheme for Manga. This could mean that Manga can be treated as a kind of "continuities" which are used as scripts in film productions, sometimes called storyboards. However, we also confirmed that we should revise the scheme to describe the connections between frames, balloons, and texts. For example, it is impossible to encode all of the correlation of abstract units and images, like a section and multiple page images, in a text data structure,

4 Reformatting Systems

The Library of Congress in the US started a reformatting program, whose aim is mainly the preservation of the appearance of physical books or documents[11]. We plan to make reformatting systems, but the objective is different from L.C.'s. Our system will be designed for converting physical forms into digital forms, and then back to physical formats like: books to scrolls, scrolls to books, books to books, and so on. The idea comes from discovery and observation as follows. The extant *Ouma Jirushi*s, the early color-printed publications in Japan, are all in scroll formats, except for one in Tsurumi University, which is in a set of folios in which holes for binding remain. All the materials contain the number of leaves and volumes, printed in the center of each folio. It would mean that the *Ouma Jirushi* was originally binded in or printed for a book form, but later was re-formed into a scroll format. This kind of shift in formats can now be observed in printed music. Players sometimes reform a sheet or book of music to a long sheet or scroll of music for making easy to follow notes without turning over any pages or sheets. For narratives which contain time-sensitive content, scrolls would be more suitable than books. A format style would not have developed from being a primitive one to being an advanced one in history. Digitally reformatting systems will provide opportunities to verify the propositions and re-examine which formats are the most suitable for the content that existing materials have.



```
<text>
  <body>
    <div type="section" n="s1">
      <div type="page" n="p4" corresp="sp4">
        <div type="frame" n="f1">
          <fs><f name="area" fVal="4034 3108 6076 5112"></f></fs>
          <div type="balloon" n="b1"><fs><f name="area" fVal="5284 3144 5932 3900"/></fs></div>
        </div>
        <div type="frame" n="f2">
          <fs><f name="area" fVal="2387 3108 4756 5112"></f></fs>
          <div type="balloon" n="b1"><fs><f name="area" fVal="4354 3156 4696 4112"></f></fs></div>
          <div type="balloon" n="b2"><fs><f name="area" fVal="2549 3156 3442 4092"></f></fs></div>
        </div>
        <div type="frame" n="f3">
          <fs><f name="area" fVal="563 3108 2327 5112"></f></fs>
          <div type="balloon" n="b1"><fs><f name="area" fVal="1739 3144 2273 4008"></f></fs></div>
          <div type="balloon" n="b2"><fs><f name="area" fVal="635 3156 1379 4212"></f></fs></div>
        </div>
```

Figure 2

5 Reformatting Manga from Recto First to Verso First Reading Style

According to a news story in *the Yomiuri Shimbun*[13], French translation version of *Pink*, which is a title of Japanese Manga, was published with two formats; one for reading from left to right pages, which is an ordinary way to Western culture, and one for reading from right to left pages, which is an authentic Manga style. The latter is easy to make a translated version, however the sales have not been good as compared to the former. This kind of trials has also been carried out in US[12]. Making a verso first reading style version costs publishers much time and money for re-fabrication. For example, let's compare an original format in Fig.1 to a recto-first format in Fig.2.

Since there is no change in images, the directions of frames and balloons are the same on both formats. However, the directions of language reading is different from them. Then, editors have tried to make recto-first formats into verso-first formats for English versions. Fig.3 is a verso-first format with two ways of changing: translational and reflectional symmetry movements. The first row in Fig.3 is a mirror image of originals. Each frame on the second row in Fig.3 is results of translational symmetry movement. Then, on the first row a balloon direction is the same as a frame direction, but is not on the second row. On the other hand, information pictured in frames is kept on the second row instead. Which way is adopted in editing translation versions depends on editors' expertise. The aim of our system is to support the re-formatting processes.

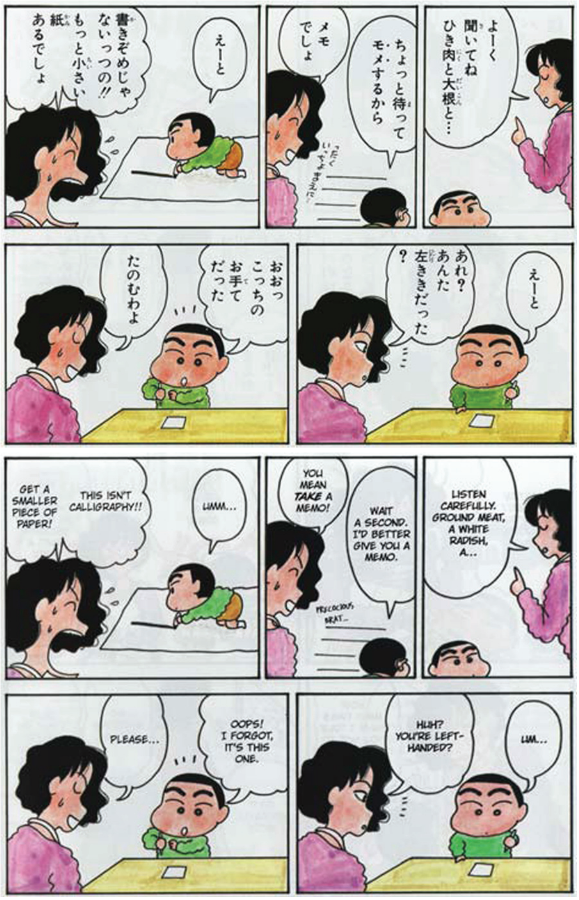

よーく 聞いてね ひき肉と大根と…
ちょっと待って モ・メするから
メモでしょ
ったく いっちょまえに
えーと
書きぞめじゃないっつの!! もっと小さい紙あるでしょ
えーと
あれ? あんた左ききだった?
…
おおっ こっちのお手てだった
たのむわよ
LISTEN CAREFULLY. GROUND MEAT, A WHITE RADISH, A...
WAIT A SECOND. I'D BETTER GIVE YOU A MEMO.
YOU MEAN TAKE A MEMO!
PRECOCIOUS BRAT...
UMM...
THIS ISN'T CALLIGRAPHY!!
GET A SMALLER PIECE OF PAPER!
UM...
HUH? YOU'RE LEFT-HANDED?
OOPS! I FORGOT, IT'S THIS ONE.
PLEASE...

Fig.1 and Fig.2

6 Collation of Image Units

In order to make the reformatting system, first of all, we have to lay down a scheme for collation of graphic resources. The results from the scheme will be used as data for analysis to find out heuristic rules for automatic frame relocation, then for instructions in converting systems. Making a scheme for collation of graphic units will be a new challenging target. Now we are experimenting with descriptions to find ideal schemes for collation of graphic units.

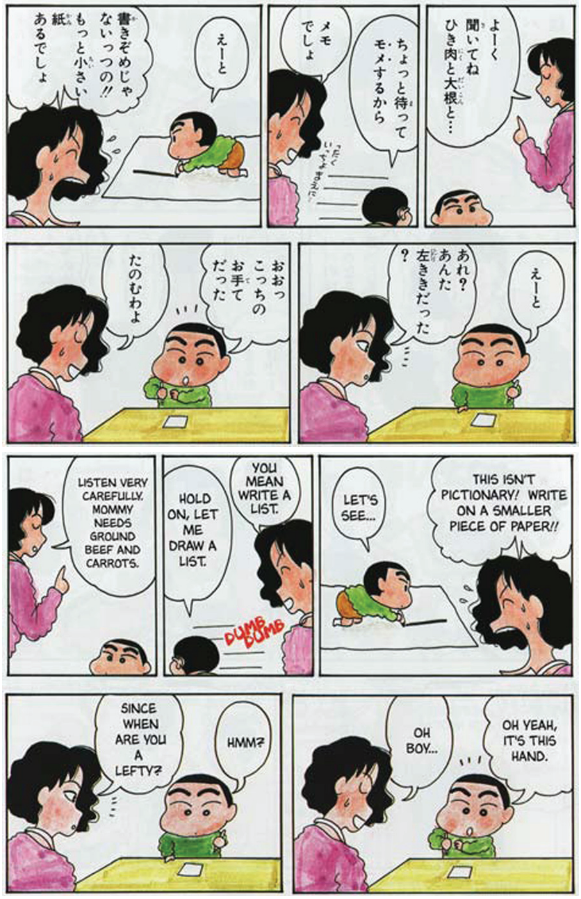
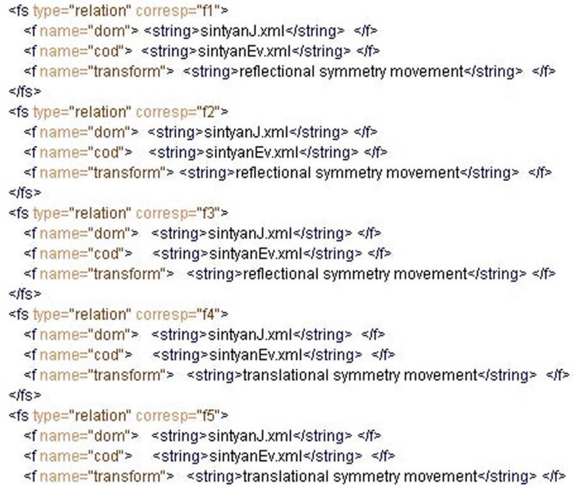

よーく聞いてね ひき肉と大根と…
ちょっと待って モメするから
メモでしょ
えーと
書きぞめじゃないっつの!! もっと小さい紙あるでしょ
えーと
あれ？ あんた左ききだった？
おおっ こっちのお手てだった
たのむわよ
LISTEN VERY CAREFULLY. MOMMY NEEDS GROUND BEEF AND CARROTS.
HOLD ON, LET ME DRAW A LIST.
YOU MEAN WRITE A LIST.
DUMB DUMB
LET'S SEE...
THIS ISN'T PICTIONARY! WRITE ON A SMALLER PIECE OF PAPER!!
SINCE WHEN ARE YOU A LEFTY?
HMM?
OH BOY...
OH YEAH, IT'S THIS HAND.

Fig.1 and Fig.3



```
<fs type="relation" corresp="f1">
   <f name="dom"> <string>sintyanJ.xml</string>  </f>
   <f name="cod">  <string>sintyanEv.xml</string> </f>
   <f name="transform">  <string>reflectional symmetry movement</string>  </f>
</fs>
<fs type="relation" corresp="f2">
   <f name="dom">  <string>sintyanJ.xml</string> </f>
   <f name="cod">    <string>sintyanEv.xml</string> </f>
   <f name="transform"> <string>reflectional symmetry movement</string>  </f>
</fs>
<fs type="relation" corresp="f3">
   <f name="dom">   <string>sintyanJ.xml</string> </f>
   <f name="cod">     <string>sintyanEv.xml</string> </f>
   <f name="transform">   <string>reflectional symmetry movement</string> </f>
</fs>
<fs type="relation" corresp="f4">
   <f name="dom">   <string>sintyanJ.xml</string>  </f>
   <f name="cod">     <string>sintyanEv.xml</string>  </f>
   <f name="transform">   <string>translational symmetry movement</string>  </f>
</fs>
<fs type="relation" corresp="f5">
   <f name="dom">   <string>sintyanJ.xml</string> </f>
   <f name="cod">     <string>sintyanEv.xml</string>  </f>
   <f name="transform">   <string>translational symmetry movement</string> </f>
```

Sample Code

For example, encapsulating balloon information into frames in image structures looks to make easy to specify types of movements in reformatting. However, it is still a problem in order to make adequate results to what extent we should encode instructions about image movements in collation data on the assumption that image processors can adjust fluctuating location information encoded in XML data.

7 Conclusion

Now we are making markup texts of several titles of Manga following the TEI scheme as much as we can, and seeking a reference scheme for collation of image unites in comparing a Japanese version with the two different formats of English translation versions. We are planning to alter the TEI scheme to match our requirements especially making useful collation data for reformatting systems.

8 References

[1]**L.Burnard and S.Bauman eds.** (2007) *The TEI Guidelines P5*, TEI

[2]**K.Ohya and S.Tutiya** (1999) "Links between link elements in compound data units" in Japanese, *IPSJ SIG Technical Report* Vol.99. No.48., IPSJ

[3]**K.Ohya** (2006) "Markup problems: Syntactical analysis and steps to their resolution" in Japanese, *TEI Day in Kyoto 2006 Report*, Kyoto University

[4]**K.Ohya** (2008) "Management of links between link elements to represent correlation on link structures" in Japanese, *IPSJ SIG Technical Report* Vol.2008. No.100., IPSJ

[5]**K.Yasuko et al.** (2006) *Witchblade Takeru* Vol.1 in Japanese, Akita Shoten

[6]**K.Yasuko et al.** (2007) *Witchblade Takeru Manga* Vol.1, Bandai Entertainment.

[7]**K.Yasuko et al.** (2008) *Witchblade Takeru Manga* Vol.1, Top Cow Productions

[8]**Y.Usui** (1992) *Kureyon Shinchan* Vol.1 in Japanese, Futabashya

[9]**Y.Usui** (2002) *Crayon Shinchan* Vol.1, Comicsone

[10]**Y.Usui** (2008) *Crayon Shinchan* Vol.1, DC Comics

[11]**The Library of Congress** (2006) "Preservation Digital Reformatting Program", The Library of Congress(USA)

[12](2006-11-21) "'Witchblade' Manga in Two For-

mats", ICv2.com

[13](2007-04-03), "Page order makes an impact on sales" in Japanese, Yomiuri Shimbun

Text and Pictures in Japanese Historical Documents

Takaaki Okamoto
Ritsumeikan University
04c0004@sch.otani.ac.jp

Majoring in Japanese History, I have been conducting research with its focus on handwriting in historical documents. Through searching for documents with the same handwriting from a great number of historical documents, I collate documents written by the same person, analyzing them with questions of not only what is written, but also why this specific person wrote this specific document, and why this particular text is included in this specific textual body.

Since comparisons and analyses of handwritings reveal what text-only analyses cannot, history and palaeography have been arguing importance of the study of this kind. Despite this argument, however, we have only seen insufficient progress in the study of handwriting. A major reason for this derives from the fact that researchers have not been facilitated with a suitable environment to help them conduct ever repeating tasks of not only searching the same letters and characters in enormous amounts of texts of enormous amounts of documents, but also comparing them. This paper proposes a computer system to facilitate such an environment for historians and its possible applications for pictorial materials. Please note that what I am aiming for here is not automatic identification of handwriting by computer. Rather, the computer is to help researchers' identification by organizing, searching, and displaying data.

In order to identify somebody's handwriting, first I have to search letters and characters that appear both in a 'standard' document written by that person and a document to be compared with. Since this is a fairly exhaustive research to look for all comparable letters and characters, I use computers to sort out and organize information about what kind of letters are located in what place in which document, and based on this, I have been doing research on methods to search and display characters or strings of characters.

There are two kinds of information about where a character is located in a document. One is logical information, expressing information about the location by using terms such as page, line, and column. You can find examples of this kind in book indexes and the computer's full text search. The other is to point out the location visually. This is as if someone brings a book with him, opening the page, and pointing out to you exactly the place you are searching.

In the system I propose here, I separate characters in the text from each other, putting them into relational database in which each character is treated as one record. For each character, I assign these two kinds of locus information. One is its location in the logical structure of the text—what number in terms of the page, the line, and the column; and the other is the physical location as expressed in the coordinates on the digital image. This kind of assignment, manually conducted, results in not only reconstructing text by assembling such characters, but also specifying the location of a character in a digital image of the document. When character data and image data are linked through coordinates, we can create a character catalogue by cutting characters out of document images. We can also search for a character or combination of those and highlight them in the digital image, as if someone brings a book with him, opening the page, and pointing out to you the exact character you want as Fig. 1 shows.

I belong to the Japanese Culture Research Group a part of Global COE (Center of Excellence) program Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University, and the Group puts focus more on *ukiyo-e* and other visual material than on textual ones such as archival documents. For this reason, I am now working on to systematize information on 'what image is where in what material' in digitalized images that the same university's Art Research Center has been ever accumulating. This system works like putting tags with some notes on pictures, but by using computerized tags rather than paper ones. So, what is the point of using the computer here?

First of all, among the many merits of this procedure, we can create other contents, based on every piece of information about what is where. For instance, if we place a mark on the publisher's seal in an *ukiyo-e* print and input its data, we can not only search for and display it in the database, but also make a program that creates a list of the publisher's seals by cutting out the image parts with the publisher's seals and generate it in PDF or other formats upon creating a layout that links the data inputs.

Secondly, using the computerized tags means that we can show data on what is where in what material by using URLs. Imagine the situation, in which one researcher may want to inform another researcher of a part of a picture in the collection of the Art Research Center archives. He might attach the whole image or only a part of it to his mail with detailed explanation. Instead of such toils, when utilizing the data on the web, he would only need to create the data on 'what is where in what material' and send the URL. The receiver would then access the URL from his browser and inspect the picture with a tag attached to it and read the notes.

Untill now computers have mainly been used in the humanities as a means to create databases. Starting with catalogues, now we can examine both full texts and images of textual materials. While the catalogues and full texts, and the catalogues and images, are respectively linked, we have seen little progress in the linkage between the full texts (or, in the case of pictorial materials, data on various elements of the picture) and the images.

Since this system makes this possible and people can use it personally, researchers can organize their research materials of full texts and images which they have collected by linking them. Besides such personal use, this system can be developed to be a system for multiple users. I believe that by systematizing data on 'what is where in what material,' we can suggest further possibilities of applying computers for the humanities, and that makes significant contribution to not only study of handwriting, history and paleography but also the humanities in general.

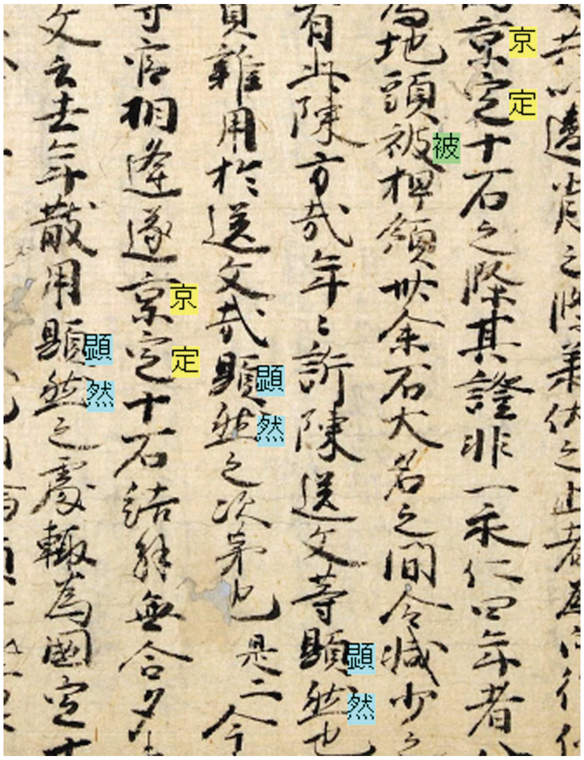


Fig.1

Text-Image Linking Environment (TILE)

Dorothy Carr Porter
Digital Humanities Observatory
dot.porter@gmail.com

Doug Reside
Uniersity of Maryland, College Park
dresied@umd.edu

John Walsh
Indiana University
jawalsh@indiana.edu

Introduction

To create the next generation of the technical infrastructure supporting image-based editions and electronic archives of humanities content, we are developing a new web-based image markup tool, the Text-Image Linking Environment (TILE), through a collaboration of the Maryland Institute for Technology in the Humanities, Indiana University Bloomington, the Royal Irish Academy, the University of Oregon, and Harvard's Center for Hellenic Studies. Despite the proliferation of image-based editions and archives, the linking of images and textual information remains a slow and frustrating process for editors and curators. TILE, built on the existing code of the AXE image tagger, will dramatically increase the ease and efficiency of this work. TILE will be interoperable with other popular tools (including both the Image Markup Tool and the Edition Production and Presentation Technology suite) and capable of producing TEI-compliant XML for linking image to text. We will also put the image linking features of the newest version of the Text Encoding Standard (TEI P5) through its first rigorous, "real world" test, and, at the close of the project, expect to provide the TEI with a list of suggestions for improving the standard to make it more robust and effective. TILE will be developed and thoroughly tested with the assistance of our project partners, who represent some of today's most exciting image-based editions projects, in order to create a tool generated by the community, for the community, with the expectation that, unlike so many other tools, it will be used by the community.

History of Images in the Digital Environment

Texts, from the earliest classical inscriptions to most twentieth-century correspondence, were originally inscribed on such physical objects as stones, papyrus scrolls, codex manuscripts, printed books, and handwritten and typewritten letters. As editors transfer a text from its original inscription, some of this context is necessarily obscured. Further, editors must often make potentially questionable decisions as they interpret the unclear or damaged text on the original artifact. A good editor will, of course, highlight such interventions in textual notes, but such notes, usually in small type and inconveniently separated from the main text, often go unread. The inclusion of page facsimiles can make the editorial process more transparent, but in print editions the reproduction of multiple, high quality images is often prohibitively expensive. Digital facsimile editions, on the other hand, may be distributed far less expensively, and so many editors are now choosing to publish their facsimile editions online.

The growth of the Internet as a public space in the early 1990s led to the first generation of widely-accessible scholarly electronic archives, and even at this early stage many projects integrated images into their work in significant ways. The Valley of the Shadow (1993), provided images for some of letters in the collection (in relatively low resolution), and the Rossetti, Dickinson, and William Blake Archives brought together encoded texts and images or parallel viewing and study.[1] The relationship between image and texts in these archives is quite simple: for example, the page image of the source of the edited text in the *Valley of the Shadow* or the *Rossetti Archive* may be opened in a separate window, but the links go no deeper than the page level. One cannot, for instance, link from a word in the edited text to its location in the image or click an interesting area in the image to read an annotation.

At the same time that these relatively open-ended online archives were under development, other scholars were taking advantage of digital technologies to build self-contained scholarly editions. Some of the earliest efforts include the *Wife of Bath's Prologue on CDROM* (Chaucer 1996), the *Electronic Beowulf* (Kiernan 1999), and the *Piers Plowman Electronic Archive*, Vol. 1, (Langland 2000). As with the online archives, these early editions were limited in how closely they linked image and text. The *Electronic Beowulf* did provide some annotations linked to areas on the manuscript folio image, but there are few of these as the coordinates for each had to be added to the HTML "by hand."

As the community of scholars developing image-based projects has grown in the past decade, tools have been created that are actively used for project development. As of November 14, 2008, the project investigators know of no fewer than ten tools or collections of tools that allow users to edit or display images within the context of textual projects or editions. These range from those that simply display an image alongside a text, to very robust software suites which support the development of complete image-based projects with substantial functionality beyond simple text-to-image mapping.

The simplest tools enable the viewing of images alongside text transcription, either for editing or for display. Juxta, developed through Networked Infrastructure for Nineteenth-century Electronic Scholarship (NINES) <http://www.nines.org/tools/juxta.html> provides a window for viewing image files (if provided) alongside transcriptions, which could be very useful for an editor checking readings or adding annotations, but does not provide any method for connecting the image with the text beyond the page level. Similar is the Versioning Machine, developed by Susan Schreibman at the University of Maryland Libraries (http://v-machine.org/): a display tool for comparing encoded texts that also enables page images to be linked to the text at the page level. These tools are both useful, but for those scholars who seek to include more fine-grained linking in their projects they are not suitable.

There are also tools that support the linking of image to transcription or annotation. The Edition Production and Presentation Technology (EPPT), developed by Kevin Kiernan at the University of Kentucky under the aegis of the Electronic Boethius and ARCHway projects (http://www.eppt.org/eppt/) is a set of tools that have been developed in and run through the Eclipse software development platform. One of the main functions of the tool is to link transcription to an image of text, although it provides much more robust functionality. The Image Markup Tool (IMT), under development by Martin Holmes at the University of Victoria, BC, is the first tool to output complete and valid TEI P5 XML. The IMT enables a user to place a series of annotations on an image, resulting in a file that validates against the regular (unmodified) TEI P5 schema, and then enables the user to create HTML for the display of those annotations online. The IMT is very simple and easy to use, and is in many ways a model of the type of tool that we will be developing in this project - it does one thing, and it does it very well. Unfortunately, the IMT runs only on Windows machines and cannot be easily ported into new web-based projects. TILE will interoperate with the constrained IMT TEI format.

There have also been some efforts to build tools to automate the creation of links between transcribed text and image of that text. Hugh Cayless at UNC-Chapel Hill has recently developed a system for automating im-text linking, a process he presented at the Text Encoding Initiative Member's Meeting, November 2008,[2] and Reside has also developed the Word Linking tool, originally developed for the Shakespeare Quartos project.

The Ajax XML Encoder (AXE), also developed by Reside, allows users with limited technical knowledge to add metadata to text, image, video, and audio files. Users can collaboratively tag a text in TEI, associate XML with time stamps in video or audio files, and mark off regions of an image to be linked to external metadata. At present the web-based image tagger allows users to select regions in an image and store the coordinates of this region in a database, but it does not provide tools to make use of this data once it is stored. The text tagger allows a user to specify a relaxNG schema and then tag a text using this schema, but it requires users to enter coordinates for image links by hand (it does not, at present, interface easily with the image tagger). The tools in AXE were always intended to be interoperable and to have the functionality described in this narrative, and this current collaboration allows us to move the suite to the next stage of its development.

The Tool

TILE will be based primarily on the Ajax XML Encoder (AXE). Through TILE, we will extend the functionality of AXE to allow the following:

- Semi-automated creation of links between transcriptions and images of the materials from which the transcriptions were made. Using a form of optical character recognition, our software will recognize words in a page image and link them to a preexisting textual transcription. These links can then be checked, and if need be adjusted, by a human.

- Annotation of any area of an image selected by the user with a controlled vocabulary (for example, the tool can be adjusted to allow only the annotations "damaged" or "illegible").

- Application of editorial annotations to any area of an image.

- Support linking for non-horizontal, non-rectangular areas of source images.

- Creation of links between different, non-contiguous areas of primary source images. For example:

    - captions and illustrations;

    - illustrations and textual descriptions;

- analogous texts across different manuscripts

We are especially concerned with making our tool available for integration into many different types of project environments, and we will therefore work to make the system requirements for TILE as minimal and as generic as possible.

Notes

[1] Valley of the Shadow: Two Communities in the American Civil War, Virginia Center for Digital History, University of Virginia (http://valley.vcdh.virginia.edu/); The Complete Writings and Pictures of Dante Gabriel Rossetti, A Hypermedia Archive, edited by Jerome J. McGann, University of Virginia (http://www.rossettiarchive.org/); Dickinson Electronic Archives, edited by Martha Nell Smith, Online. Institute for Advanced Technology in the Humanities (IATH), University of Virginia (http://www.emilydickinson.org/); The William Blake Archive. Ed. Morris Eaves, Robert N. Essick, and Joseph Viscomi. (http://www.blakearchive.org/).

[2] Hugh Cayless, "Experiments in Automated Linking of TEI Transcripts to Manuscript Images," presented at the Text Encoding Initiative Member's Meeting, November 2008. http://www.cch.kcl.ac.uk/cocoon/tei2008/programme/abstracts/abstract-166.html

References

Carlquist, J. 2004. "Medieval Manuscripts, Hypertext and Reading. Visions of Digital Editions. Literary & Linguistic Computing 19.1,105-118.

Chaucer, G. 1996. *Wife of Bath's Prologue on CDROM*. Edited by P. Robinson. Cambridge University Press.

Holmes, M. 2007. Image Markup Tool v. 1.7. [http://www.tapor.uvic.ca/~mholmes/image_markup/] Accessed 2008-11-13.

Kiernan, K.. 2005. "Digital Facsimiles in Editing: Some Guidelines for Editors of Image-based Scholarly Editions." *Electronic Textual Editing*. Ed. Lou Burnard, , Katherine O'Brien O'Keeffe and John Unsworth. New York: Modern Language Association, 2005. [preprint at http://www.tei-c.org/About/Archive_new/ETE/Preview/kiernan.xml]

Kiernan, K. 1999. *The Electronic Beowulf*. University of Michigan Press.

Kirschenbaum, M. G. 2002. Editor's Introduction: Image-based Humanities Computing. *Computers and the Humanities* 36.1, 3-6.

Langland, W. 2000. *Piers Plowman Electronic Archive*, Vol. 1. Edited by R. Adams. University of Michigan Press.

TEI Consortium, eds. 2007. "Digital Facsimiles." Guidelines for Electronic Text Encoding and Interchange. [Last modifieddate:2008-07-04].[http://www.tei-c.org/release/doc/tei-p5- doc/en/html/PH.html] Accessed 2008-07-25.

Authorship Attribution, The Large and Small Effect Sizes of Divergence as Classification

Mike Ryan
Duquesne University
michaelryan@acm.org

Patrick Juola
Duquesne University
juola@mathcs.duq.edu

A common method of performing authorship attribution (or text classification in general) involves embedding the documents in a high-dimensional feature space and calculating similarity judgments in the form of numeric "distances" between them. Using (for example) a k-nearest neighbor algorithm, an unknown document can be assigned to the "closest" (in similarity or distance) group of reference documents. However, the word "distance" is ill-defined and can be implemented conceptually in many different ways. We examine the implications of one broad category of "distances".

This notion of distance can be generalized to dissimilarity judgements without previous embedding in a space. An example of this is the Kullback-Leibler Divergence which calculates an information-theoretic dissimilarity measure, effect size, between two event streams where the events are not necessarily independent and thus cannot be directly tabulated as simple histograms. This kind of "distance" can easily be incorporated into a text classification system.

To a topologist, a "distance" is a numeric function D(x,y) between two points or objects, such that

- D(x,y) is always nonnegative, and always positive if x != y
- D(x,y) = D(y,x)
- D(x,y) + D(y,z) >= D(x,z)

However, there are many useful distance-like measures (technically known as "divergences") that do not have all these properties. In particular, divergences such as the Kullback-Leibler divergence and vocabulary overlap are not the same when measured from different basis. So, assume that you have two documents A and B and want to find the divergence between them. The methods would be to find the divergence of B from A, D(A, B) or to find the divergence of A from B D(B,A). These two ways of applying the same divergence will give you different results.

The two results being different is important because it implies that there is different information captured by each one. This has been shown to be true for some divergence measures (Juola Ryan, 2008) but upon reviewing those previous results, it seems there is rarely a case where information spread across both divergences i.e. the information gained from one is better then an average of the two. The problem then becomes that we do not know which one will contain more information. Because of this we have come up with criteria for selecting one over the other. The criteria devised are simple; we will either use the max of the two methods, or the min of the two methods.

To test this, we are in the process of applying several divergence functions to a standardized corpus [the AAAC corpus (Juola, 2004)] of authorship attribution problems using the JGAAP framework (Juola et al, this conference). We will compare each run normally, backwards, then both taking the max and min.

Preliminary results using the Kullback-Leibler Divergence and the LZW Distance indicate that using the max of the two divergences will on some problems increase accuracy by up to four-fold. We plan to continue this work using other divergences; if this finding continues to hold, we consider this to be an important step to eliminating some of the "ad-hoc-ness" of the current state of authorship attribution, as we will be able to provide some steps to analyzing not merely what methods perform best, but what extensions of these methods can be used to improve their performance.

ArchInSite: Augmented (Reality) Architecture

Eric Sauda
UNC Charlotte
ejsauda@uncc.edu

Nick Ault
UNC Charlotte
nwault@uncc.edu

Zac Porter
UNC Charlotte
ztporter@uncc.edu

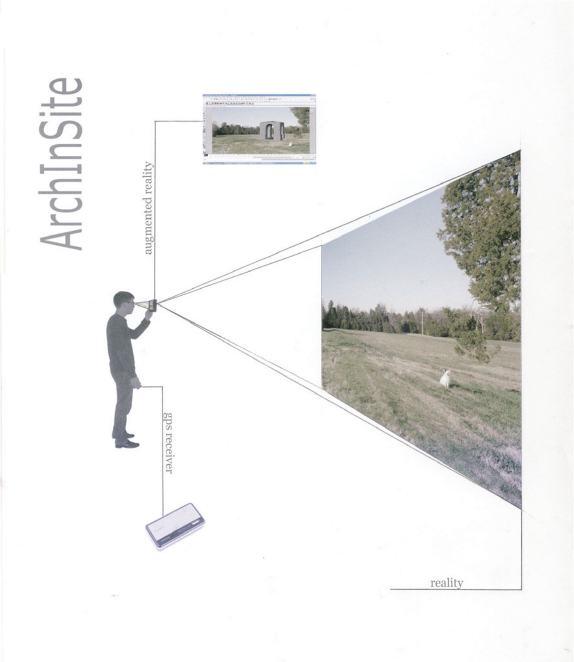


Fig. 1 Diagram of ArchInSite operation.

ArchInSite is an augmented reality application that combines three-dimensional modelling, video compositing and a global positioning system into a handheld device. This system allows a user to move through the landscape, while using a mobile device to view virtual models accurately placed in space. ArchInSite is unique in its combination of a hand held device with GPS and video compositing. This system can be useful for visualizing architectural designs, sculpture and augmented environments directly on real site, while simultaneously interacting with and making design decisions within the virtual environment. Further, ArchInSite offers the possibility of creating a virtual world woven into real space, both as a means of displaying architectural information and exploring design alternatives within the existing conditions. At the conceptual level, privileging the perspectival view challenges the conventional place of orthographic projections as the media of architectural design.

Architectural Background

While the vast majority of the architectural community is concerned with scaled, planometric drawings that represented spaces from a God's eye view, we are interested in exploring a technology that would allow us to visualize buildings in a real-time, interactive first-person perspective. Our reasoning is simple enough—if one experiences the world in a first-person perspective, why not design buildings in a first-person perspective? It appears to us that too often architects are referencing non-experiential spatial concepts, such as a geometrical relationship between adjacent rooms that is only evident in plan, instead of the direct experiential nature of inhabiting a building. We consider first-person perspectives of a building to be indubitable knowledge for architectural experience. Likewise, we consider an overall, spatial plan of a building to be comparable to the philosopher Edmund Husserl's description of a belief. Therefore, a progression through a building, in which one is able to take in several first-person perspectives (indubitable knowledge), allows one to begin making some basic assumptions (beliefs) about the overall spatial plan of the building. So, while the experience of a building progresses from first-person perspectives to an overall, spatial plan, the standard practice for designing a building progresses in reverse, from planometric drawing to perspective. We see this problem as an opportunity to create a device that would provide the kinds of real-time, perspectival visualizations necessary to design in a first-person perspective.

It is our contention that the analysis of these first-person perspectives, not as our own perspectives, but as an ideal, archetypal perspective, should be the driving force of an architectural design. Like Husserl, we believe that a further consideration of things in and of themselves can provide a foundation for more conceptual and abstract thoughts.

Background/ Related Work

It is possible to trace the concept of computer generated reality to Ivan Sutherland's pioneering work with head mounted displays and his idea of a window into a virtual world (Sutherland). Much of this early work focused on the creation of a world separate from our 'normal' world, which became known as virtual reality. The term 'augmented reality' was first introduced by two Boeing engineers (Caudell), who created the Wire Bundle Assembly Project, which combined a heads-up, see-through, head-mounted display that registered computer produced diagrams superimposed on real world objects using half silvered mirrors and position tracking. Steven Feiner's work on 'A Touring Machine' (Feiner) combined bulky GPS equipment with optical overlay heads-up display to geo-spatially register and display text information about campus building interactively. Ongoing work by Behzadam and Kamat (Behzadam) at the University of Michigan is using video compositing and GPS data to feed a head mounted video display for possible use in construction.

Research Issues and Implementation

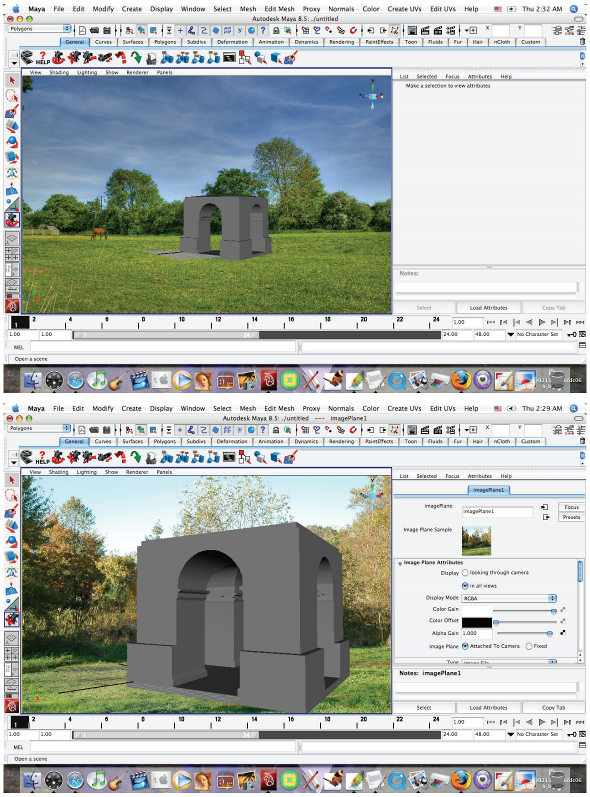
Augmenting the sensible world with additional information is, on it's face, a simple and reasonable goal. Because we never just look without intention, the ability to enrich the visual context with information makes possible new forms of understanding and interaction.

Augmented reality has raised important ideas about both hardware and software. The hardware issues have largely revolved around issues of display and registration (by contrast, generating 3d models and placing them on still photographic backgrounds is well understood). Display choices are largely focused on heads-up display technologies, with choices between see-through displays and video compositing. Registration of the model image with the real world is a difficult and on-going problem, requiring specialized equipment (head tracking accelerometers) and limited environments (position tracking). While these efforts continue to show promise, their current state is still largely fragile and more suitable to research labs than to wide spread use.

We have developed a prototype for a system that incorporates three-dimensional modelling, GPS tracking, an internal camera and video compositing to produce a device that creates a hybrid condition between the virtual and real worlds. The current implementation consists of a Sony UX micro pc running Maya modelling software, a GPS data parser and video capture software. The system utilizes a MEL script that enables the built in camera and the data streamed from the GPS to reorient the model in virtual space based upon the position of the user. A system such as this is feasible for virtually all camera/GPS enabled phones (currently estimated at 400 million units/year worldwide); this makes their use potentially ubiquitous.

Results

We have successfully tested the use of ArchInSite for viewing of proposed architectural models in the landscape. The system is readily understood and embraced by users, and while the registration problems endemic with head mounted systems do not completely disappear, separating the view port from direct attachment to users (by substituting hand held display for heads-up display) provides much more flexibility for the users to comprehend the registration of the virtual and the real. We began with preselected models that were placed at specific points in the landscape; we then implemented (using Maya's default modelling tool set) the ability to make changes to the model while concurrently viewing virtual architecture integrated into its site.


Maya File Edit Modify Create Display Window Select Mesh Edit Mesh Proxy Normals Color Create UVs Edit UVs Help
Thu 2:32 AM
Autodesk Maya 8.5: ../untitled
Polygons
General Curves Surfaces Polygons Subdivs Deformation Animation Dynamics Rendering PaintEffects Toon Fluids Fur Hair nCloth Custom
View Shading Lighting Show Renderer Panels
List Selected Focus Attributes Help
Notes:
Select Load Attributes Copy Tab
No Character Set
MEL
Thu 2:29 AM
imagePlane1
Focus
Presets
Notes: imagePlane1

Make a selection to view attributes

imagePlane: imagePlane1

Image Plane Sample

Image Plane Attributes

Display ○ looking through camera ● in all views

Display Mode RGBA

Color Gain

Color Offset

Alpha Gain 1.000

Image Plane ● Attached To Camera ○ Fixed

Open a scene

Fig. 3 Screenshots showing change of user position and corresponding change in model and composited background.

We have three major findings.

1. The system that we have implemented utilizes eas-

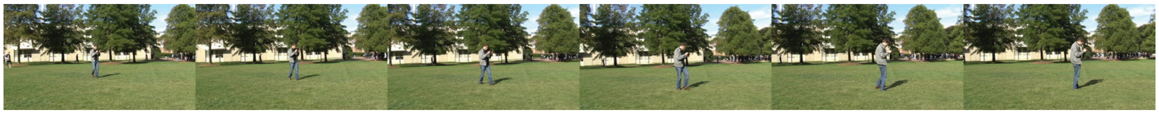


Fig. 2 ArchInSite user changing position on an outdoor site.

ily obtainable handheld devices rather than heads-up displays, and side steps the difficulty of solving registration problems. Our users have found this handheld device approach to be a convincing window that they can connect directly and viscerally to the context.

2. We believe that ArchInSite can prove valuable as a design tool for architects and designers. The ability to see work with a model on site is useful because it gives you a richer and more fulsome understanding of the context than is possible when pursued through a distance by sketches, renderings and/or drawings. Coupling this visual understanding with the ability to alter the computerized model allows the designer to sensitively recalibrate and redesign the model in the actual setting.

3. We think that significant opportunity exists to construe a virtual world that is threaded inside the real world and supplements it in interesting ways. Architecture was often used as a classical rhetorical device for memorizing facts and telling stories and we see the potential for its re-emergence threaded through the real world.

Further Work

We are currently developing a new modelling and compositing implementation that will use custom software rather than off-the-shelf applications that often have high levels of system overhead. Our further goal is to move this application, or a variation of it, over to a more compact and ubiquitous device, such as the iPhone, a device that currently contains the hardware necessary to drive the software.

We will also be teaching an architectural studio course that will use ArchInSite as the primary means of design, seeking to challenge the hegemony of the orthographic.

References

Behzadan A.H., and Kamat V.R. **(2005), Visualization of Construction Graphics in Outdoor Augmented Reality, Proceedings of the 2005 Winter Simulation Conference, Orlando, FL**

Caudell, Thomas P., and Mizell, David W. (1992). *Augmented Reality: An Application of heads-Up Display Technology to Manual Manufacturing Processes*, 1968 Fall Joint Computer Conference, AFIPS Conference Proceeding **33**, 757-764 (1992).

S. Feiner et al., **"A Touring Machine: Prototyping 3D Mobile Augmented Reality Systems for Exploring the Urban Environment," Proc. 1st Int"l Symp. Wearable Computers (ISWC "97), IEEE CS Press, Los Alamitos, Calif., 1997, pp. 74-81.**

Sutherland, Ivan E. (1968). *A Head Mounted Three Dimensional Display*, 1968 Fall Joint Computer Conference, AFIPS Conference Proceeding **33**, 757-764 (1968).

An Image-Based Document Reader with Editing Functions for Education and Research on Digital Humanities.

Hiroyuki Sekiguchi
Ritsumeikan University
h-seki@fc.ritsumei.ac.jp

This paper focuses on an electronic document reader we have developed, a new method of lecture using this reader, and some possible examples of its utility for research on digital humanities.

Introduction

Recently, more and more documents have been stored in digital archives. Electronic documents have many advantages, e.g., high performance when searching text, instantaneous document transmission, and negating the need for archive space of the corresponding physical documents. Because of this, we can read thousands of digitized books anywhere with our notebook PCs.

Despite these benefits, however, it is often said that most people find it difficult to remember the contents of digital documents when reading them on a computer display, compared with reading the physical paper document. We can propose several reasons for this. Firstly, any physical movements which indirectly assist our recollection, such as eye-movement or page turning, are not experienced when reading digital documents. Secondly, making notes or memos or marking on the electronic document is usually not possible, and even if it is, we cannot do it as easily as on paper.

Some reports say that taking memos or highlighting parts of documents improves our understanding and remembering. Since such activity requires proactive decisions on selecting contexts, these actions activate our brain much more than silent reading. Of course, these markings and memos help us recollect our thoughts during a later reading of the document.

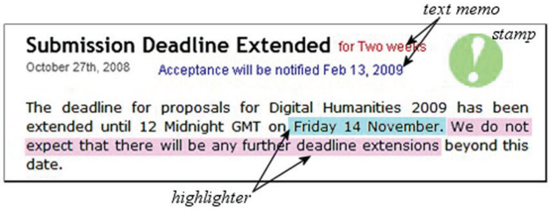
To overcome these shortcomings, we have developed a document reader that enables us to freely mark digitized documents. The reader we have developed has several marking functions, such as, highlighting areas of text, inserting bookmark stamps, and adding memos to the text. One of the most unique features of our reader is to realize these mark-up functions with usability almost equal to that of paper media. This point is important because such an easy-to-use text reader has not prevailed yet.

Our reader stores mark-up data in a file separate to the original document. As the mark-up data is written in a simple text-numeric format, it can be easily used by a lot of software applications, such as spreadsheets, statistics analysis or image processing packages.

In this paper, we will start with an explanation of our reader's functions and then introduce two applications using it: a new method of lecturing using the reader, and examples of using mark-up data for the study of digital humanities.

Functions Of Our Reader

We now introduce editing functions equipped in our document-reader. An example text containing mark-up highlights, stamps, and memos written on the text is shown in Figure 1.



Submission Deadline Extended for Two weeks

October 27th, 2008 Acceptance will be notified Feb 13, 2009

The deadline for proposals for Digital Humanities 2009 has been extended until 12 Midnight GMT on Friday 14 November. We do not expect that there will be any further deadline extensions beyond this date.

Figure 1: A sample document with markings

Highlights:

We can highlight text just by click-and-dragging the mouse along the text. Any other operations like opening menus or clicking icons are not required. Dragging from left-to-right highlights the text in red, while, dragging from right-to-left highlights in blue. You may think that highlighting in blue is a little bit difficult due to the right-to-left movement, but in exchange for this inconvenience, you do not have to choose a separate color to use. That is, blue is naturally used for short key words or important contexts.

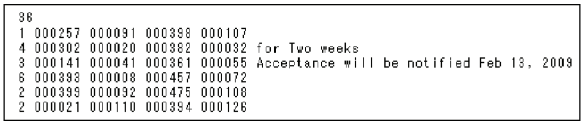
Stamps:

If you push the space key, a stamp appears on the text. The stamp has two faces: "?" and "!". The intended purpose of the stamp mark "?" is to mark hard-to-understand concepts, or points that need to be followed up later. The "!" mark is used for "interesting" or essential points of concepts.

Stamps also act as bookmarks. If you push the "<" or ">" key, the reader jumps to the previous or next stamp, respectively. If there is no following stamp in the current document, the reader jumps to the next bookmark in the following document.

Text memos:
You can add notes or memos anywhere in the document. Push the “Tab” key to enter text-input mode. Characters you type on your keyboard will appear at the position of the mouse cursor. To exit input mode, push the “Tab” key again. Unlike the other editing applications, our reader does not show any text-input-box which spoils the look and feel of writing on the paper.

User’s mark-up data file format:
A user’s mark-up data is stored in an associated file. An example of a user’s mark-up data is shown in Figure 2. You can see that it is quite a simple form of text-numeric data. The number in the top line indicates the position of the cursor when closing the document on the previous reading. Each of the following lines describes the marker type and the coordinates of the surrounding rectangle. In the case of a text-memo, the user’s text is added as a character string after the coordinate values. These data are used not only for displaying mark-ups, but for other various purposes, such as taking mark-up statistics, extracting highlighted-regions from the document, and comparing comments written by different reviewers. We will now show some examples below.



```
36
1 000257 000091 000398 000107
4 000302 000020 000382 000032 for Two weeks
3 000141 000041 000361 000055 Acceptance will be notified Feb 13, 2009
6 000393 000008 000457 000072
2 000399 000092 000475 000108
2 000021 000110 000394 000126
```

Figure 2: Mark-up data of Figure 1

Use During Lectures

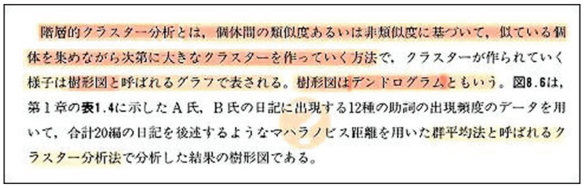
Thanks to the improvement of our computer literacy and network technology, computer-assisted education, (so called e-learning), has become more and more popular. Our university has several computer rooms, but they are mainly used for learning computer software or programming. So, we have been developing a new style of lecturing using our reader to make the most use of computer rooms for the classes of humanities. The arrangement of a typical class is as follows.

First, the students read the lecture notes for the next class with our reader, highlighting or putting stamps on phrases they cannot understand clearly.

Next, the teacher receives student mark-up data via the network prior to the lecture. The lecturer checks the student reports with a summarized view (shown in Figure 3), and then, he examines which areas of the lecture students are having a hard time understanding. As a result, the teacher can modify his lecture so that it is more suitable and understandable for the students.

There are several other benefits for the students, too; 1) They can understand lessons more deeply by paying attention to areas they could not clearly understand. 2) Their sense of participation is promoted because their mark-up results directly affect the lecture content or explanation.

These major benefits are achieved by relatively simple editing actions by the students. According to a class evaluation taken at the end of the semester, most of our students approved of this style of lecture. By investigating the mark-up data gathered from them, we can extract difficult-to-understand keywords and sections of the lecture. In the near future, it may be possible to find relations between their marking style and their learning results.



階層的クラスター分析とは，個体間の類似度あるいは非類似度に基づいて，似ている個
体を集めながら次第に大きなクラスターを作っていく方法で，クラスターが作られていく
様子は樹形図と呼ばれるグラフで表される。樹形図はデンドログラムともいう。図8.6は，
第1章の表1.4に示したＡ氏，Ｂ氏の日記に出現する12種の助詞の出現頻度のデータを用
いて，合計20編の日記を後述するようなマハラノビス距離を用いた群平均法と呼ばれるク
ラスター分析法で分析した結果の樹形図である。

Figure 3: Summarized views of student marks.

Using For The Study Of Digital Humanities

A particular feature that paper media have is that we can easily write on them, however, the reality is that, it is usually prohibited to write on research materials. By using our reader, we are completely free from such restrictions. This is one of the benefits of our writable reader used as a researching tool.

In addition, by displaying marks or comments written by more than one researcher on the same document, we can clearly recognize which points should be focused on, and what the differences are between their ideas. This function is also useful for referee reading or proofreading performed by several reviewers.

Finally, since each piece of mark-up data indicates its position in the material, it is easy to extract marked regions from the original material. After gathering these regions, they can be shown as listings or KWIC (keyword in context) forms. We can also use the marked region as a template for searching over the document. In this way, this reader can be used as a front-end to researching tools.

Conclusions

In this paper, we have confirmed the utility values of our digital document reader for understanding computerized documents and its usefulness in education. As mentioned above, our reader can be used in several ways for the research of digital humanities. Especially, by handling materials such as image documents, our reader is suitable for handwritten documents and those with a large number of figures.

You can download our text reader from our website. We would like you to try it and give us feedback. Please feel free to stop by our poster session, as we are going to demonstrate it and introduce new methods for research on digital humanities.

URL: http://www.img.is.ritsumei.ac.jp/~h-seki/beeReader/

New Digital Tools at the William Blake Archive

William Shaw
UNC-Chapel Hill
wsshaw@email.unc.edu

In this talk, I plan to discuss and demonstrate my recent work at the William Blake Archive—namely, my development of a virtual lightbox application. I will discuss this work in the context of our editorial rationale, software roadmap (including our user interface redesign, currently scheduled for completion in early 2009), and our past approaches to digital tool development.

The user interface of the Blake Archive currently relies on Java-based tools developed at the Institute for Advanced Technology in the Humanities (IATH), University of Virginia, in the mid-1990s. These applets, ImageSizer and Inote, allow users to view some of Blake's work at true size and to read the Archive editors' image descriptions. They have proven useful and durable, but we have long sought to incorporate their functionality into a more powerful tool: a virtual lightbox application that provides users with image manipulation capabilities, features for object annotation, and the ability to collect and compare Blake's work across genres, media, and time periods.

As technical editor of the Blake Archive, I have spent the past several months developing this software, and the Archive has successfully deployed it on our testing site. My poster will discuss the ways in which this software greatly expands the functionality of the Blake Archive in particular, and, in general, the ways in which it can be adapted to any digital project concerned with imaging in various contexts, with a special emphasis on art-historical and manuscript studies.

In terms of the Blake Archive, The primary goal of the lightbox is to give serious Blake scholars an indispensible tool for comparative analysis. By allowing users to collect different plates from different copies of an illuminated book (for example) and view them all at true size, the lightbox provides a fundamental tool of art-historical analysis that has hitherto been absent from the Blake Archive (as well as from similar projects, such as the Rossetti Archive). In addition, its annotation features allow users to explore image descriptions and editorial commentary; they also enable users to search currently loaded images by keyword, motif, or other markup characteristics.

Furthermore, its backward-compatibility with SGML Inote annotations permits a smooth transition between previous annotation/description schemes and the new, XML-enabled Lightbox environment.

In more general terms, I will discuss the architecture of the Blake Archive lightbox, emphasizing its open-source (MIT/X11) licensing and simple API. Its design, which is intended to be both straightforward and flexible, allows it to be incorporated into any web-based digital project with ease. Its simple, XML-based image annotation format is adaptable to a wide range of image markup needs. Finally, its installation procedure—which relies on simple JavaScript functions—will ensure that it is easy to integrate into other projects.

In addition to discussing the features and development of the lightbox, I will explain how it fits into the editorial rationale of the Blake Archive. As editors Morris Eaves, Robert N. Essick, and Joseph Viscomi have pointed out, the priority that we grant to the media, methods, and histories of artistic production has dictated a feature of the Archive that influences virtually every aspect of it. It is utterly fundamental: we *emphasize the physical object—the plate, page, or canvas—over the logical textual unit—the poem or other work abstracted from its physical medium*. This emphasis coincides with our archival as well as with our editorial objectives. (Eaves, Essick, and Viscomi, “Principles”).

This art-historical emphasis on the physical object, rather than exclusively on the literary text, is continued in the lightbox. It not only allows users to focus on the physical objects themselves, rather than editorial apparatus, but also complements both our editorial principles of diplomatic transcription and our archival principles of size fidelity and comparative analysis.

In conclusion, I plan to discuss my development of the lightbox application; to explain how the application fits into both the Blake Archive site redesign and, in theory, answers the needs of other projects; and to argue that, as a tool, it is both an important part of our user experience and a manifestation of the editorial principles that have guided the entire history of the William Blake Archive.

Works Cited

Eaves, Morris, Robert Essick, and Joseph Viscomi. “Editorial Principles: Methodology and Standards in the Blake Archive.” *The William Blake Archive*. April 15, 2005 <http://www.blakearchive.org/blake/public/about/principles/index.html>

-----. “Technical Summary of the William Blake Archive.” *The William Blake Archive*. 16 June 2008 <http://www.blakearchive.org/blake/public/about/tech/index.html>

TADA Research Evaluation Exchange: Winning 2008 Submissions

Stéfan Sinclair
McMaster University
sgsinclair@gmail.com

Winners: Dave Beavan, Susan Brown, J. Stephen Downie, Carlos Fiorentino, Patrick Juola, Shelly Lukon, Peter Organisciak, Geoffrey Rockwell, Susan Schreibman, Kirsten Uszkalo

In the spring of 2008 the Text Analysis Developers' Alliance organized a digital humanities tools competition called T-REX, modelled on similar competitions such as MIREX (music information retrieval) and TREC (text retrieval competition), (cf. Downie 2006). The community response to T-REX was very positive and among the many submissions received, judges selected winners from the following categories:

- Best New Web-based Tool
- Best New Idea for a Web-based Tool
- Best New Idea for Improving a Current Web-Based Tool
- Best New Idea for Improving the Interface of the TAPoR Portal
- Best Experiment of Text Analysis Using High Performance Computing

The categories above deliberately cover not only working tools, but also ideas, designs and preliminary experiments; a primary objective of T-REX is to encourage the involvement and collaboration of programmers, designers, and users. More information on the categories, the judges, and other aspects of T-REX are available at http://tada.mcmaster.ca/trex/.

The organizers and winning participants of T-REX would like to propose a cluster of posters that showcase various aspects of the research done. In particular, we will prepare seven "half" posters presenting relevant aspects from each of the winning T-REX submissions. In addition, an eighth "half poster" will provide an overview of the inaugural TREX competition, lessons learned, and new initiatives for the second round. Below are brief descriptions of each of the seven project posters.

Susan Brown et al.,
Degrees of Connection Tool (New Tool)

This linkage tracing tool allows users working on a large collection of documents to explore the linkages within the collection based on the semantic tags it contains. Our prototype based on the Orlando textbase traces links between people mentioned in different XML documents based on co-occurrences of a small set of key tags that occur across many documents: personal names, organization names, places, and titles. This exploits the tagging to get at connections between people that may not be made by direct linkages between documents, but rather by the co-occurrence of tags within two documents, or a pathway from document x to document y by way of document z in which different tags common to x and y occur in z. It is a way of getting at implicit but nevertheless potentially important linkages, and while it emerges in this case from an interest in literary history, the tool could be useful to other fields ranging from journalism to creative writing, sociology or psychology. It provides a new way of exploring the large digital collections researchers are increasingly using. The poster will 1) outline the principles on which the prototype is based, 2) list key features for a fully-developed, generalized version, 3) explain our application of graph theory to the tagged text, and 4) outline the design challenges that emerged in the development of the prototype. A live demo will allow attendees to test the prototype.

David Beavan,
Collocate Cloud (Idea for Existing Tool Improvement)

Clouds of information e.g. keywords, tags or words, are a very useful way to aggregate and present vast quantities of data. These clouds have gone on to be used in many web 2.0 sites. As such they are becoming a well known and understood visualisation by many users.

TAPoRware currently provides a Word Cloud visualisation, which shows the frequency of words in a document. Scholars often wish to go further, to see how a particular word is used, by examining which words co-occur near their search word. TAPoRware already has this Collocation tool, showing the results in tabular form.

The Collocate Cloud would merge the collocation output and the cloud visualisation technology. It will show the collocates of a particular search word in cloud format. The alphabetical ordering of the Collocate Cloud would allow the user to find or discount a word quickly. Frequencies and collocational strength are shown by size and brightness, letting these terms stand out visually.

Carlos Fiorentino, The Magic Circle (Idea for New Tool)

The Magic Circle is an information glyph that allows scholars to visually summarize combinations of the lexical information included in text collections (typically frequency data about words, lemmas, and parts of speech) and the bibliographic information attached to these texts. The glyph consists of a set of rings organized outwards from the centre and divided in wedges or sections. The lexical data determine the size of the centre, which also shows a word, a lemma, or a part of speech, and the total number of search results for that word, lemma, or part of speech within a specified work or set of works. The bibliographic data is related to authors specified by the user, and the rings allow the user to analyze how the search results are distributed in different collections as well as in different periods of time. The color sets of the rings follow patterns of associations with variations in hue, tone and saturation. A comparative scale helps the user to understand the volume of information found in context with the whole volume of information present in the collections.

Alejandro Giacometti et al., Ripper Browser (New Tool)

The Ripper Browser is a prototype for rich-prospect browsing of text collections. Rich-prospect browsing interfaces are designed to aid research tasks such as exploration and synthesis by providing both a meaningful representation of each item in a collection and tools to manipulate their visual organization (Ruecker 2003). The Ripper browser offers an environment for exploration and interaction with digital text documents. The system creates tiles that contain faceted information about each document. The tiles can be manipulated with a series of controls to reveal or hide details, organize them according to a particular hierarchy, or select a specific group. By adapting the size of the tiles, the Ripper browser allows researchers to visualize the complete collection and the precise information they need about each document in view at all times. The Ripper browser was developed in web-native technologies: HTML, JavaScript, and uses the jQuery library. It is configured to use text collections provided by the MONK Project. The Ripper browser is part of an ongoing effort to understand the potential of rich-prospect browsing and improve on our strategies for designing rich-prospect tools. It has allowed us to experiment further with meaningful representation, increased our understanding of the importance of sequences, and provided insight into new possibilities for organization in visualizations.

Patrick Juola & Shelly Lukon, Back-of-the-Book Index Generation (Experiment in HPC)

This is actually a work-in-progress; as we have detailed elsewhere (Juola, 2005, ACH/ALLC; Lukon and Juola, 2006, DH2006), we are working on a program to apply standard ML techniques, including latent semantic analysis, to the problem of back-ofthe-book index generation. LSA implicitly uses huge document-by-term matrices to determine which words appear in similar contexts and are therefore good candidates for grouping under a single index term.

The sheer size of this matrix makes it difficult to work without HPC; one of the tools we are using is the 200+ node Beowulf cluster available at Duquesne University Computer Science Department. We analyze the document to be indexed (which can in theory be arbitrarily large but in practice will be about novel-sized) to select candidate words (mostly nouns, via POS tagging) for indexing, then use LSA to identify potential relationships among those words.

Peter Organisciak, Bookmarklet for Immediate Text Analysis (Improving TAPoR)

This idea is of an online interface for the generation of TAPoR bookmarklets on demand.

Bookmarklets are browser bookmarks that run javascript code. They provide value to text analysis tools in two way: ease and ubiquity. They allow one-click connection of content to tool and, more importantly, allow it to be run on whatever content the user is at.

One problem of bookmarklets is that they are static, which means that customization of the query is limited. One solution would be to call up an interface every time the bookmarklet is called. Doing so, however, is an impediment to the core concept of ease. Rather, through an interface for creating customized bookmarklets, a user can create single-purpose bookmarklet buttons that do the same command every time, immediately and directly.

Kirsten C. Uszkalo, Throwing Bones (Idea for New Tool)

The Throwing Bones interface operates as a means to discover meaningful relationships within a corpus of texts. These relationships will appear as a series of piles, which the user can zoom into and out of, shuffle through, and examine closer for more comprehensive, annotated information. For example, in the case of a corpus of early modern witchcraft trials, a user might want to see the relationship between animal familiars and accusers. After shuffling, the top item in a pile would show the number of familiars, while the cards beneath show the number of accusers, illustrating a connection between the imagination of accusers and the presence of familiars in trials and texts. The piles could also be based on geographic, temporal, textual, or relationship proximity. The concept behind Throwing Bones is that the interface will not only offer the pleasure of play, but also erudite and serendipitous textual analysis.

Bringing Southern Oral Stories Online

Natasha Smith
University of North Carolina, Chapel Hill
nsmith@email.unc.edu

Joshua Berkov
School of Communication Arts
jberkov@gmail.com

Cliff Dyer
University of North Carolina, Chapel Hill
jcd@unc.edu

Hugh Cayless
New York University
hugh.cayless@nyu.edu

In recent years, oral histories have become an alternative medium for interrogating the past and have assumed a prominent place in historical inquiry. They offer unique perspectives from individuals who have witnessed history in the making and often yield unparalleled insight into the lives and times that they record. Long constrained by the media used to record them, oral histories are increasingly the target of digitization initiatives that seek to preserve these voices and make them heard by disparate audiences. “Oral histories of the American South” (http://docsouth.unc.edu/sohp/) is just such an initiative.

From its beginnings as a pilot project in 2004, this endeavor, funded by the Institute of Museum and Library Services, grew rapidly and attracted attention. Documenting the American South, a digital publishing program at the Carolina Digital Library and Archives (CDLA), worked in close cooperation with a number of departments at the University of North Carolina at Chapel Hill – the UNC Library, Southern Oral Histories program (SOHP), and School of Education – and applied new technologies, open standards and some of the tested practices highlighted in other DocSouth collections. Far from being simply a collection of digitized documents, these oral histories undergo rigorous analysis by subject specialists. The practice of applying such scholarship has added considerable value to other recently published collections and has brought together the perspectives of historians and the first-hand experiences of witnesses to history. Finally, it is an experimental project to build an interface to simultaneously display audio and transcripts from interviews.

The project will be complete by the time of the conference in June 2009 and we are proposing to present a poster on the work we have done, highlighting the challenges and joy experienced by the Project team of librarians, humanists, technologists, and – not to forget – users who participated in several usability testings and studies.

The process of creating digital documents of these oral histories is indeed a challenge. From the beginning—the process of selection is a daunting prospect unto itself. It is the task of historians from the SOHP to select 500+ representative interviews from a collection that now numbers well over 4,000. This includes careful attention to privacy and copyright issues. Only interviews free of restriction are considered for inclusion in the project collection.

Cassette tapes and typescripts are the raw materials from which these interviews are remade into digital objects. Audio engineers at the Southern Folklife Collection, using the best available hardware, software, and their own considerable expertise, produce digital audio files in both WAV (preservation) and MP3 (access) formats. These audio data are of the highest possible quality, and comply with international standards and best practices for the creation and preservation of digital audio content. The typescripts are all encoded in TEI P4 (with the plan of conversion to P5), conforming to level 4. But the story doesn't end here…

Once created, these newly digitized interviews are subjected to intense historical analysis by subject specialists in the Southern Oral History Program (SOHP). With the guidance of scholarly advisors, specialists, mainly PhD History students, read through the transcript, write abstracts, create descriptive titles, and select particularly powerful segments. Their decisions are based on a number of criteria, from intrigue to major historical relevance and from uniqueness to conformation of commonality. Once these segments are chosen, the PhD students then assign keywords (category/subcategory combinations) to each of these segments, and a given segment will usually have multiple keywords. The keywords are then given a rank, depending on their relevance to the segment. A given segment might have 3 keywords, all of varying degree if importance or relevance to the segment itself. Assigning these keywords and then prioritizing them is what makes our soon-to-be released new advanced search so effective. A user can type in a keyword and immediately retrieve the interviews that are most relevant due to our efforts to assign and then prioritize these keywords. Finally, PhD students write short descriptions for the selected segments and provide the historical context to the interviews.

Once complete, these interviews return to DocSouth, where they are prepared for publication. DocSouth staff collect various existing metadata from a number of sources to create rich records for each interview, enhancing the XML transcripts and adding the interviews to the MySQL database. These metadata are subsequently used by library catalogers to generate MARC records to further enhance retrieval. A trained specialist listens to the interviews and inserts timestamps into the files based on the selections made by the SOHP. These text timestamps become points of entry into the audio—clicking on them plays back the audio for that segment.

The interviews are displayed as flat HTML files, generated in advance from the TEI source files using XSLT. We have implemented a cutting-edge advanced search interface that is constructed with Python/Django forms and templates. The form framework provides hooks for robust input validation, while the templates separate the content from the display for us, so non-tech-savvy designers can help craft an elegant search interface. The search index is built on Solr, a lucene-based search engine which allows for fast, flexible searching on full-text or fielded queries. Database access is in many places facilitated by a robust Object Relational Mapper written in Python called SQLAlchemy. This allows queries to the DB to occur seamlessly, without having to rely on cumbersome, fragile SQL queries.

The success of bringing these valuable documents out of archives and into the eyes and ears of the public come from the efforts of its many participants, including librarians, historians, and education specialists. The project team members and the various institutions and interests they represent work together to produce digitized oral history interviews that are the result of the application of technology solutions, adherence to standards, scholarship, and exceptional technical expertise. All of these efforts bring additional value to oral history interviews that are moving and insightful stories of an American South in the process of profound and irrevocable transformation.

References

http://www.tei-c.org/wiki/index.php/TEI_in_Libraries:_Guidelines_for_Best_Practices

http://www.djangoproject.com/

http://www.sqlalchemy.org/

http://lucene.apache.org/solr/

A Historical GIS Analysis of the Landscape Compositions: A Case Study of Folding Screens *"Rakuchu-Rakugai-zu"*

Akihiro Tsukamoto
Ritsumeikan University
atv28073@fc.ritsumei.ac.jp

Introduction

Recently, the use of Geographic Information Systems (GIS) to analyze historical space has attracted interests; this approach is known as Historical GIS (HGIS) a subdiscipline in Digital Humanities (William and Thomas, 2004). In this paper, I propose a new methodology to analyze compositions of landscape in paintings with GIS, which I apply for analyzing Japanese screen paintings known as *rakuchu-rakugai-zu* (洛中洛外図), created between the 16th and 18th centuries. *Rakuchu-rakugai-zu* provides contemporary bird's-eye views of the city and environs of Kyoto, Japan's capital at the time. The paintings capture man-made structures such as residences and palaces of prominent samurai and court nobles, temples and shrines; natural features such as hills and rivers; and festivals and other human activities. Nearly all of these works consist of a pair of folding screens of six panels each. At present, over 100 such sets are known to have survived.

Because the scenes of Kyoto painted on these screens provide historical information not found in written texts, they have attracted interests of scholars in various fields. Particularly noteworthy among these studies are the following three results of detailed analyses of space on *rakuchu-rakugai-zu* screens. Takeda (1966) classifies these screens into three types, according to their era of creation and drawing method. (i) The 'standard' type, commonly known as the '*first-generation*,' depicts scenes from the late 16th century, mainly views of urban districts of both Shimogyo, i.e., the southern half of the town, and Kamigyo, i.e., the northern half of the town, with the surrounding hills as a backdrop. (ii) The '*variant*' type offers close-up depictions of specific subjects or districts. (iii) The 'developed' type, commonly known as the '*second-generation*,' depicts scenes from the 17th century, particularly of Nijo Castle and downtown Kyoto, with hills in the background. Takeda argues that the changes in drawing method reflect a shift from seasonal nature paintings employing elements of the 'famous views' painting tradition to genre paintings reflecting trends of the times.

The knowledge obtained from the previous research concerning changes in the subject matter and geographic range of *rakuchu-rakugai-zu* is essential to our understanding of the screen compositions. However, such previous research has not shed light on how actual views of Kyoto were transferred to the restricted dimensions of the screen surface, nor on how real geographic data were manipulated in the process of making that transfer. This is because the past analyses have relied solely on the simple mapping of landmarks. To analyze the compositional methods employed to paint these scenes, not enough is simply mapping landmarks; we must also determine which landmarks and districts were distorted, and how. In this paper, I propose a new approach, HGIS, which offers an analytic methodology utilizing GIS to measure distortions in the landscape depicted on the screen. This methodology visualizes distortions in the drawn space by linking the positions of landmarks as they appear on the screen painting and on survey maps, and transforming the configuration of the screen accordingly. Combining the results obtained by this method with those of previous research should provide us with more detailed and precise understanding of the drawn space on these screens.

Methodology

The methodology I introduce here uses GIS spatial analysis functions to scan the screen surface onto a survey coordinate grid based on the relative positions of landmarks in the screen painting. The analytical procedures go as follows (Fig. 1):

1. Derive coordinate values for landmarks, both on the screen and on a survey coordinate grid;

2. Generate a link table from these two point-data sets;

3. Use projective transformation and rubber sheeting techniques to project the screen surface onto the survey coordinate grid; and

4. Project the areas of the rubber sheet-derived polygons onto the screen.

This process gives visual representation to differences between actual space and the space drawn on the screen. Results show that screens painted in the 17th century and later distorted actual space more than screens painted in the 16th century, indicating a decrease in adherence to perspective-like conventions. This trend toward greater distortion suggests a shift in landscape drawing methods away from realism. Increasing deformation may be attributed to changes in political regime as well as an expansion of the public's geographic awareness. I see significant advantages of using GIS in understanding *Drawn Area* and *Regularity in Drawing*, which I will explain in detail here.

Drawn Area

Using GIS visualizes a drawn area exactly, which means we could know which landmarks are included or not in the painting. Depending upon the area included, we could know how much of forced inclusion, meaning spatial distortion, occurs. In the first-generation screens, for example, the projective transformation-derived polygons and rubber sheeting-derived polygons are similar in shape. The second-generation screens, however, reveal a pronounced tendency to forcibly incorporate landmarks to the north and south of the city, even though this results in various expedient changes in the drawn area of the left screen and a progressive narrowing of the background scenery in the right screen. Finally, the variant-type screens may be explainable as the ones with their focus on specific districts, for which the forced inclusion of landmarks as employed in the second-generation screens proved insufficient. This proves distortion progresses over time.

Regularity in Drawing

Using GIS makes it possible for me to analyze regularity in drawing, i.e., the locations of spatial abbreviations and exaggerations on the screens. First-generation screens exhibit orderly increases in area value, suggesting that specific conventions similar to the rules of perspective were followed to achieve the precise geographic positioning of landmarks. The result is a realistic rendering of geographic space. In contrast, second-generation and variant-type screens exhibit numerous instances of distorted space, and area value increases variously from locale to locale. This suggests that valued less was the actual positional relationships of landmarks in the drawing of these scenes. This means that GIS clearly shows that in terms of their rendering of real-space positional relationships the first-generation screens are superior to these later ones.

Conclusions

Until now, there has been little research applying GIS to analyze landscape drawing methods. This case study shows, however, such an analysis is feasible when applied to paintings covering a wide area of the city scape, such as *rakuchu-rakugai-zu* screens. The analytical methodology presented in this study involves not merely mapping landmarks that appear on the screens, but also projecting a "virtual screen" onto a survey map. I was able to implement this process for the first time by using the 'adjust,' 'area calculation,' and 'table join' functions of GIS analysis. The results obtained with this methodology offer insights from a geographic point of view into the approaches taken to space drawing and landmark positioning in these works.

Analyzing landscape drawing methods, conventional art history mentions that from the 17th century on, the depiction of politically significant landmarks became a norm and took priority over the accurate rendering of geographic locations. That is, landmarks took on the aspect of symbols, which the creators of the screens increasingly incorporated, paying less attention to their positional relationships.

In short, it turned out that my case study of *rakuchu-rakugai-zu* screens of the 16th and 17th centuries supports this point of conventional art history not from a subjective but an objective point of view as using GIS can visualize drawn area and regularity in drawing exactly, measured in quantity (Fig. 2). While this overview has provided a sense of general trends over time, I hope to follow up this study with more detailed analyses of individual screens in the context of the sociopolitical environments in which they were created.

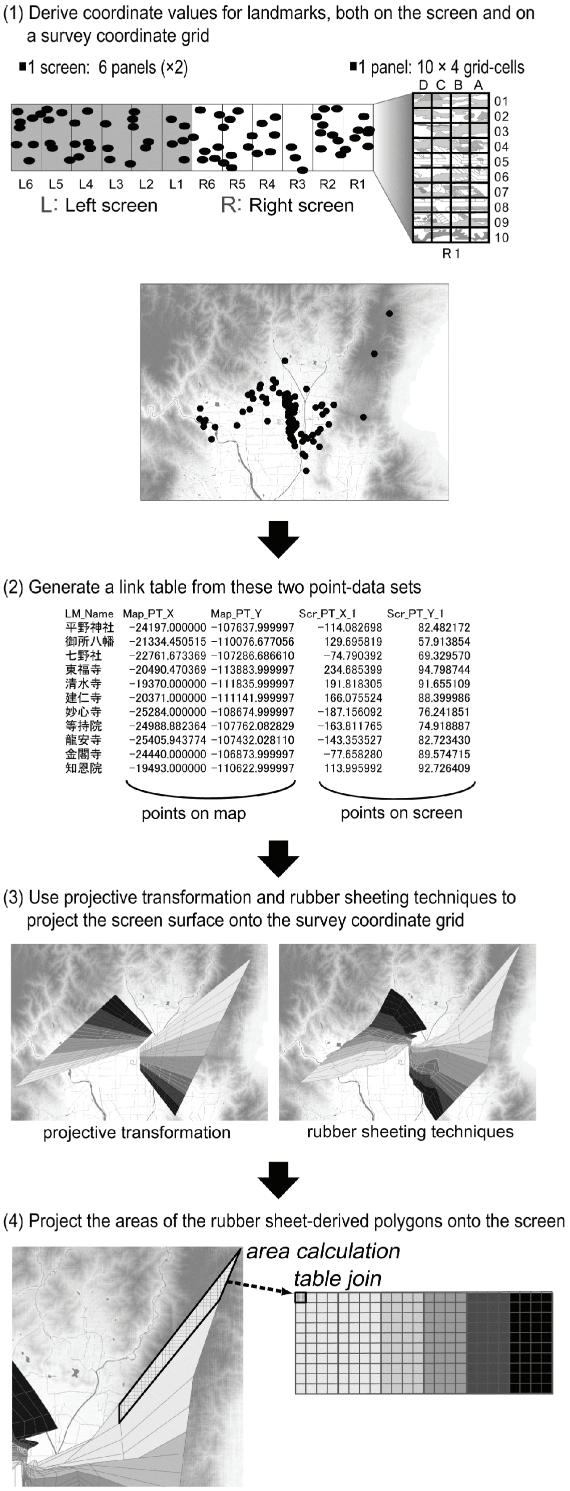
Acknowledgements

This research was conducted with assistance from the Japan Society for the Promotion of Science and the Ministry of Education, Culture, Sports, Science and Technology, Global COE Program, "Digital Humanities Center for Japanese Arts and Cultures", Ritsumeikan University.

References

Takeda, T. 武田恒夫 (1966). Rakuchu-rakugaizu to sono tenkai 洛中洛外図とその展開. In Kyoto kokuritsu hakubutsukan ed: Rakuchu-rakugai-zu 洛中洛外図, kadokawa shoten 角川書店, 1966, pp.17-32.

William, G. and Thomas, III (2004). Computing and the Historical Imagination. In Schreibman, A., Siemens, R., Siemens, R. G. and Unsworth, J. (eds), A Companion to Digital Humanities, Oxford: Blackwell Publishing, pp.56-68.



(1) Derive coordinate values for landmarks, both on the screen and on a survey coordinate grid

(2) Generate a link table from these two point-data sets

LM_Name	Map_PT_X	Map_PT_Y	Scr_PT_X_1	Scr_PT_Y_1
平野神社	−24197.000000	−107637.999997	−114.082698	82.482172
御所八幡	−21334.450515	−110076.677056	129.695819	57.913854
七野社	−22761.673369	−107286.686610	−74.790392	69.329570
東福寺	−20490.470369	−113883.999997	234.685399	94.798744
清水寺	−19370.000000	−111835.999997	191.818305	91.655109
建仁寺	−20371.000000	−111141.999997	166.075524	88.399986
妙心寺	−25284.000000	−108674.999997	−187.156092	76.241851
等持院	−24988.882364	−107762.082829	−163.811765	74.918887
龍安寺	−25405.943774	−107432.028110	−143.353527	82.723430
金閣寺	−24440.000000	−106873.999997	−77.658280	89.574715
知恩院	−19493.000000	−110622.999997	113.995992	92.726409

(3) Use projective transformation and rubber sheeting techniques to project the screen surface onto the survey coordinate grid

(4) Project the areas of the rubber sheet-derived polygons onto the screen

Figure 1. Procedures

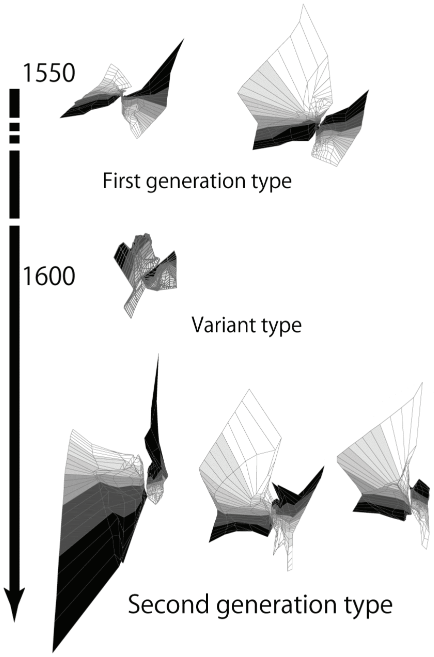


Figure 2. From Realism to Deformation

Towards an Online Edition of the Slovenian Biographical Lexicon

Petra Vide Ogrin
Slovenian Academy of Sciences and Arts, Library
petra.vide@zrc-sazu.si

Tomaž Erjavec
Jožef Stefan Institute
tomaz.erjavec@ijs.si

1. Introduction

The paper presents the project of digitization of the Slovenian Biographical Lexicon (SBL). We first describe the up-conversion of the source OCR text to a richly structured XML, encoded according to the Text Encoding Initiative Guidelines TEI P5 (TEI, 2008), using the module on biographical and prosopographical data. Next, some more challenging aspects of the conversion process are discussed, in particular the extraction of meta-data, the expansion of abbreviations into their fully inflected forms, the diachronic nature of the text, and the effects of some language technology tools developed for the Slovenian language on the efficiency of the information retrieval for Slovenian texts.

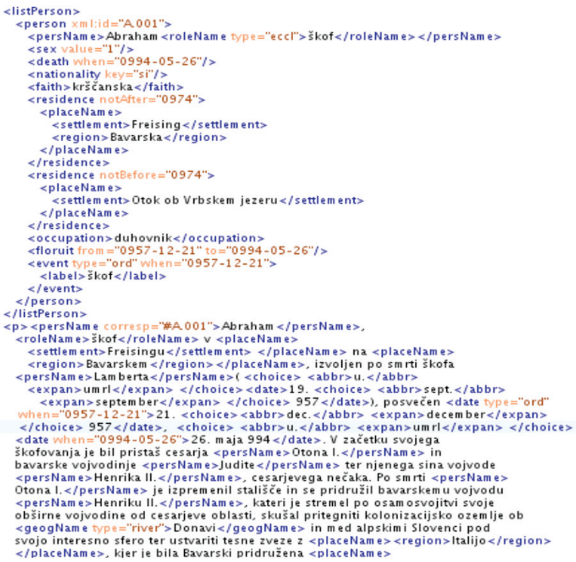
2. The SBL

The Slovenian Biographical Lexicon summarizes the lives and work of notable figures from Slovenia's cultural history. It gives a picture of Slovenia's cultural life, from its beginnings up to the contemporary time by including those who participated in its cultural development, were of Slovenian nationality or born in Slovenia, and were active in the homeland or abroad, as well as persons of foreign origin who with their work among the Slovenians influenced the Slovenian cultural life. It comprises 15 volumes plus index, with over 3,000 pages or 16 mio characters, and covers 5,031 biographical entries or, as some of the entries are family names, over 5,100 persons. Its publication spanned almost 70 years (1925-1991). SBL aims to cover not only a person's biography but also to give information about the important literature depicting their life and work or to direct a user to the whereabouts of their unpublished work, photographs, letters etc. The data in the SBL articles were always checked against the primary material source. As such, the SBL is a reliable reference for any relevant scientific research in the fields of humanities, social sciences and the history of natural sciences.

In order to widen the availability of SBL, the Slovenian Academy of Sciences and Arts (SASA)[1] and the Scientific Research Centre of the SASA (SRC SASA)[2] is undertaking the digitization of the SBL in order to make it freely available on-line, also enabling the kind of information retrieval not allowed by the nature of printed text.

3. Up-conversion and TEI P5 encoding

The SBL was first scanned and the text OCRed. The OCR was then semi-manually corrected for errors and then via a series of automatic steps converted into a rich TEI encoding. The first step, via the OpenOffice text-editor and associated XSLT TEI stylesheet,[3] converted the source text into a basic TEI structure (Erjavec and Ogrin, 2005). Since some metadata were already available in the form of an Access database, this was exported into the XML format, following the TEI P5 module on biographical and prosopographical data. Then additional metadata, such as the information for the `<floruit>` element defining the exact period for more than one occupation for a particular person in the entry, are added into the metadata structure manually. From here, metadata are added to the entries, and, in a related process, the text of the entries is further annotated.



```
<listPerson>
  <person xml:id="A.001">
    <persName>Abraham <roleName type="eccl">škof</roleName></persName>
    <sex value="1"/>
    <death when="0994-05-26"/>
    <nationality key="si"/>
    <faith>krščanska</faith>
    <residence notAfter="0974">
      <placeName>
        <settlement>Freising</settlement>
        <region>Bavarska</region>
      </placeName>
    </residence>
    <residence notBefore="0974">
      <placeName>
        <settlement>Otok ob Vrbskem jezeru</settlement>
      </placeName>
    </residence>
    <occupation>duhovnik</occupation>
    <floruit from="0957-12-21" to="0994-05-26"/>
    <event type="ord" when="0957-12-21">
      <label>škof</label>
    </event>
  </person>
</listPerson>
<p><persName corresp="#A.001">Abraham </persName>,
  <roleName>škof</roleName> v <placeName>
    <settlement>Freisingu</settlement> </placeName> na <placeName>
    <region>Bavarskem </region></placeName>, izvoljen po smrti škofa
  <persName>Lamberta</persName>( <choice> <abbr>u.</abbr>
    <expan>umrl</expan> </choice><date>19. <choice> <abbr>sept.</abbr>
      <expan>september</expan> </choice> 957</date>), posvečen <date type="ord"
 when="0957-12-21">21. <choice><abbr>dec.</abbr> <expan>december</expan>
 </choice> 957</date>,  <choice> <abbr>u.</abbr> <expan>umrl</expan> </choice>
  <date when="0994-05-26">26. maja 994</date>. V začetku svojega
  škofovanja je bil pristaš cesarja <persName>Otona I.</persName> in
  bavarske vojvodinje <persName>Judite</persName> ter njenega sina vojvode
  <persName>Henrika II.</persName>, cesarjevega nečaka. Po smrti <persName>
  Otona I.</persName> je izpremenil stališče in se pridružil bavarskemu vojvodu
  <persName>Henriku II.</persName>, kateri je stremel po osamosvojitvi svoje
  obširne vojvodine od cesarjeve oblasti, skušal pritegniti kolonizacijsko ozemlje ob
  <geogName type="river">Donavi</geogName> in med alpskimi Slovenci pod
  svojo interesno sfero ter ustvariti tesne zveze z <placeName><region>Italijo</region>
  </placeName>, kjer je bila Bavarski pridružena <placeName>
```

Figure 1: TEI-XML document excerpt with the `<listPerson>` *structure*

3. 1. The TEI structure

The encoding scheme of the SBL follows the TEI P5 Guidelines, in particular the module on biographical and prosopographical data. P5 introduces elements for a structured biographic entry, for which the information contained in the text is extracted and encoded separately using the `<person>` element, which contains information on the name(s), sex, nationality, faith, dates of birth and death, facts about residence, occupation, important life events, etc. In addition to the text, each entry of the lexicon thus also contains metadata, useful for searching and organising the SBL.

3. 2. Extracting metadata

Since some metadata not strictly connected with the person, who is the topic of the entry, such as other named persons in the entry articles, have been manually extracted from the SBL text, we are exploring the possibilities of automatic encoding to speed up the encoding process. Partially, this can be done on the basis of existing indexes of the SBL and via external resources, such as gazetteers. These, coupled with the power of regular expressions in Perl, can extract the relevant terms, such as dates, with a fairly high accuracy. A more principled solution would be to implement a general Named Entity Recognition (NER) system for Slovenian, which would recognize and categorise persons and places names, dates and other numeric expressions. Such a system does not yet exist for Slovenian, and we are exploring the possibility of writing NER rules for one of the widely used human language technology toolsets, such as GATE[4] or NooJ[5] (Bekavac, 2002).

3. 3. Abbreviations

SBL is written in encyclopaedic style, which means dense language and many abbreviations. There are several types of abbreviations: a) bibliographic abbreviations (e.g. *RDHV* for *Razprave Znanstvenega dru□tva za humanistične vede*); b) abbreviations to denote the authors of SBL articles (e.g. *R*□for *Fran Ramov*□; c) abbreviations to refer to a biographical entry within an article (e.g. *A*. for *Abraham*); d) abbreviations for certain geographic names (e.g. *Lj*. for *Ljubljana* or *Clvc* for *Celovec*); e) general abbreviations (e.g. the names of months).

While the use of abbreviations was appropriate for the printed books, it will only impair readability and make searching more difficult in the digital edition. The abbreviations are therefore expanded into their full forms and encoded with the TEI `<choice>`, `<abbr>` and `<expan>` elements. The basic resource for the expansion are lists of bibliographic abbreviations and abbreviations denoting authors of particular articles from the printed edition, and an additional abbreviation lexicon, semi-automatically compiled from the SBL.

A problem that occurs in the expansion of abbreviations stems from the fact that the Slovenian language, like all Slavic languages, is a highly inflective language. This means abbreviations have to be expanded into their appropriate full forms in the correct inflection, which depends on the context of the abbreviation. So, for example, "Rojen v Lj." (Born in Ljubljana), should be expanded to "Rojen v Ljubljani", with "Ljubljana" in the locative case. This problem is dealt with by automatic morphosyntactic tagging of the text, and then, on the basis of the tag assigned to an abbreviation, generating the appropriate inflected form of the lemma. This work uses tagging and lemmatisation models automatically obtained from annotated corpora and morphological lexicons, which have the advantage of generalising to out-of-vocabulary words (Erjavec and Džeroski, 2004).

3. 4. Language change

Another challenge regarding the language of SBL is due to its long publication period of almost 70 years. In this period the language, particularly its lexical aspects, has changed significantly. Changes affect particularly geographic names (e.g. *Curih* instead of *Zürich*) and terms denoting occupations and activities, from spelling variants to complete substitution of words. These changes will have to be taken into account to ensure adequate information retrieval. The plan is to add normalised (contemporary) forms to the forms in the text, by encoding them with the appropriate `<choice>`, `<reg>` and `<orig>` elements. The annotation will be automatic, but based on a lexicon of variants, semi-automatically compiled from the SBL. This lexicon, in itself an interesting diachronic language resource, is being built on the basis of a textual analysis of SBL, taking into account external language resources, such as dictionaries and gazetteers, e.g. of old and new names of places.

4. Access

The SBL is to be made freely accessible for full-text and structured searching. We are exploring the possibility of using the open source Apache Solr[6] search platform based on the Lucene Java search library, which enables such kinds of queries, accepts TEI / XML documents and enables different plugins and extra features, such as the lemmatization tool for Slovenian. No research has been done yet, however, on its effects upon the efficiency of the information retrieval for the Slovenian language.

References

Bekavac, Božo: Strojno obilježavanje hrvatskih tekstova – stanje i perspektive. (Computer annotation of Croatian texts – current state and perspectives) Suvremena lingvistika. 53-54 (2002), p. 173-182.

Cankar, Izidor et al. (eds) (1925-1991): Slovenski biografski leksikon. Ljubljana: Slovenska akademija znanosti in umetnosti.

Erjavec, Tomaž and Džeroski, Sašo (2004): Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words. Applied Artificial Intelligence 18 (1), p. 17-40.

Erjavec, Tomaž and Ogrin, Matija (2005): Digitalisation of Literary Heritage Using Open Standards. In: Paul Cunningham, Miriam Cunningham (eds.). Innovation and knowledge economy: issues, applications, case studies, (Information and communication technologies and the knowledge economy). Amsterdam [etc.]: IOS Press, 2005, p. 999-1006.

TEI Consortium, eds. (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Notes

[1] http://www.sazu.si/

[2] http://www.zrc-sazu.si/

[3] http://www.tei-c.org/wiki/index.php/TEI_OpenOffice_Package

[4] http://gate.ac.uk/

[5] http://www.nooj4nlp.net/

[6] http://lucene.apache.org/solr/

From the Local to the Global Sphere: Prospects of Digital Humanities for Japanese Arts and Cultures

Keiji Yano
Ritsumeikan University
yano@lt.ritsumei.ac.jp

Ryo Akama
Ritsumeikan University
rat03102@lt.ritsumei.ac.jp

Kozaburo Hachimura
Ritsumeikan University
hachimura@media.ritsumei.ac.jp

Hiromi Tanaka
Ritsumeikan University
hiromi@cv.ci.ritsumei.ac.jp

Mitsuyuki Inaba
Ritsumeikan University
inabam@sps.ritsumei.ac.jp

This poster discusses each of the research activities, issues, and prospects on Digital Humanities for Japanese arts and cultures at "the Digital Humanities Center for Japanese Arts and Cultures" of Ritsumeikan University. By doing so, we would like to suggest new directions in Digital Humanities, still fairly a new interdisciplinary research field, as well as contributions Digital Humanities could make to the humanities in general, and more specifically for the studies of Japanese arts and cultures. As we do conduct these research activities as a five-year Global COE (Center of Excellent) Program that the Art Research Center (ARC) at Ritsumeikan University has launched with the support from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) since 2007.

Our program of "the Digital Humanities Center for Japanese Arts and Cultures" is to further a study of the humanities based on Digital Humanities, taking Japanese art and culture as its subject, as centered on the historic city of Kyoto. For this purpose, we intend to make full use of the most advanced information technologies, such as digital archives, visualization, media technologies, Geographical Information Systems (GIS) and Web 2.0. By using such information technologies, we at the Center systematically organize information on Japanese art and culture, which will be open to the public and useful in a wide variety of contexts. The Center also serves as a global portal for the study of Japanese art and culture, aiming not only to create a new system of cultural exchanges among talented researchers within and out of Japan, but also to make the results of research available throughout the world. That means that we see the historic city of Kyoto as both a base for promoting scholarship on Japanese art and culture around the world, and a global hub for education and research.

Our digital archives are meant to bring together information on Japanese art and culture, both tangible and intangible, which are left unorganized all over the world. The archives of such cultural properties use information management technology to combine different types of data: text, pictures, sound, moving images and the motion of the human body. This combination of different types of date, covering both tangible and intangible cultural properties, gives us a great advantage over the conventional text-centered research. Furthermore, by systematically linking to databases on other systems, we can revolutionize the amount and quality of material available for research, leading to great improvements in the quality of research and to a new perspective on humanities studies.

For example, GIS technology will lead to further advances in the visibility and amount of data relating to the time and space content of various aspects of arts, culture and urban landscapes. Using the bi-directional format network environment of Web 2.0, we will create a portal that makes information more open, cooperative and usable. The resulting research will be compiled and distributed on-line. We will also promote research into analytical methods and archiving technology related to the exquisite cultural properties that characterize the historic cultural city of Kyoto.

We like to start this poster with a brief introduction of the program and its missions. After that, each of the following five research topics will be presented, discussing where their research interests lie in Japanese arts and cultures, and what their methodologies and challenges are in terms of digital humanities.

1. "Digital Archives for Japanese Woodblock Prints: Their Global Linkage"

2. "Virtual Kyoto: Integration of Digital Contents of Japanese Arts and Cultures on 4D-GIS with a Time Dimension"

3. “Digitization and Analysis of Traditional Dance Body Movement by using Motion Capture Technology”

4. “3D Modeling and Visualization of Japanese Traditional Arts and Cultural Assets”

5. “Collaborative Web Technologies for Japanese Arts and Cultures”

Also we will address issues and prospects we share at the Center, including the current situation and tendencies of the Humanities in Japan—urgent issues we have to tackle with our activities at the Center.

Some of the presumable issues to be addressed may include the following:

1. Humanities scholars in Japan, who tend to work domestically within their academic societies, shall partake in a global academic environment, including a Web-based one.

2. While humanities scholars tend to see the Web and information technology as mere tools, they also have to realize that using these tools may give rise to new perspectives and paradigms.

3. Japanese Humanities researchers should realize that the Internet is now advancing into the age of Web 2.0, meaning that the environment of the Web is becoming ever more crucial for humanities research. This pattern of research is common in Europe and the United States, and as a high-level institute for research and education we need to work in a similar way, in order to ensure the future development of research.

4. Not only Humanities scholars who are handling Japanese language but also other East Asian languages have been experiencing some difficulties in encoding texts or sharing textual resources on the Web. This is because an international standard of character encoding, Unicode, has limitations to represent characters in those languages.

We hope to contribute to the further discussion on Digital Humanities from the perspective of a non-Western country.

Websites:
http://www.arc.ritsumei.ac.jp/lib/GCOE/guideline_e.html
http://www.geo.lt.ritsumei.ac.jp/uv4w/frame_e.jsp

Restoring 3D Digital Woodcut Shape for Reproducing Ancient Book

Xin Yin
Ritsumeikan University
yin@cv.ci.ritsumei.ac.jp

Ryo Akama
Ritsumeikan University
rat03102@lt.ritsumei.ac.jp

Hiromi T. Tanaka
Ritsumeikan University
hiromi@cv.ci.ritsumei.ac.jp

Kazuaki Nagai
Nara University

1. Introduction

Cultural heritage is important for studying history, and some studies have been carried out to preserve it as digital data. *Hanpon* are ancient woodcut-printed books, all of which are regarded as important cultural heritage in Japan. (see Figure 1 for woodcut and *hanpon*). The woodcut looks black as its surface is covered with ink. A lot of studies about *hanpon* and woodcut have been done. However, there is little study to connect these two studies using digital techniques. The main contribution of our study is developing digital techniques for studying *hanpon* history using Woodcut information.

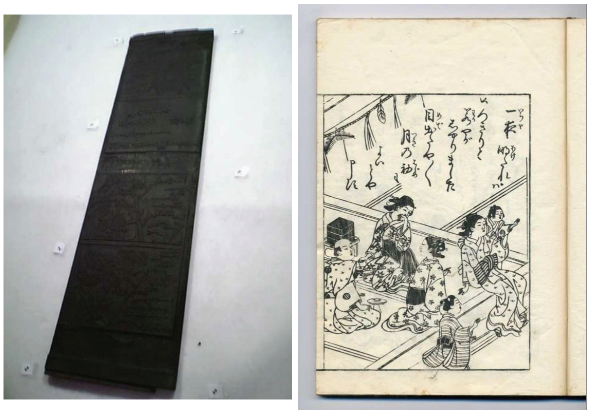


Fig. 1 Woodcut and hanpon.

There are some versions of *hanpon* whose true published time is not clear. As the paper or woodcut have shrunken over time, it is possible to guess which version is published early and which is late. Recently, some woodcuts were found, and we hope to know the original size of *hanpon* printed by these woodcuts. If we could do so, we can know which version of hanpo*n* is printed by the founded woodcuts. Warped in shape, however, they cannot be utilized to print hanpo*n* directly. To solve this problem, we propose techniques to restore woodcut shapes in a digital way so that we can reproduce ancient *hanpon*. As traditional method, a printer put some water on the back of Woodcut and the Woodcut shape can be restored near a plane. But the size of the Woodcut restored using the traditional method is different to the original one. To solve this problem, a digital technique for restoring the shape of digital Woodcut is proposed. Then we can produce original ancient hanpon.

Basically, our system is a virtual printing system. Okada[1] proposes a system for sculpturing and printing in a virtual world. Focusing on virtually sculpturing 3D objects, his system can print Japanese drawings if one paints some colors on the virtual surface of a carved object. Yet, the system uses a carved plane for virtual printing and cannot be utilized in the case of measured woodcut digital data, as real woodcuts are distorted a little. For virtual printing on the distorted surface of woodcut,[2] uses a small plane to fit to a distorted woodcut surface and make a virtual printing. However, this method ignores changes in woodcut size. To acquire a *hanpon* the same as the ancient one, it is necessary to restore woodcut shapes.

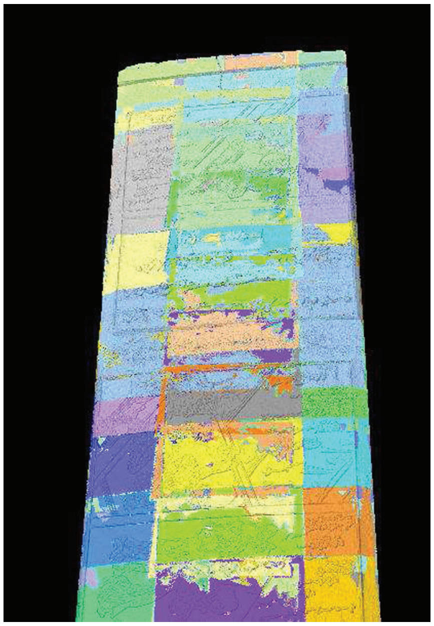


Fig. 2 Aligned 3D digital Woodcut model

For restoring woodcut shapes, we can learn a lot from some research concerned with wood drying.[3] Finite element method (FEM), for example, is utilized to simulate moist translation and shape variation. This study in the wood research field mainly focuses on how to decrease shape distortion that occurs during the drying process. As wood is an orthotropic material, we propose an algorithm for determining orthotropic direction to restore woodcut shapes.

The following is the procedure to restore the woodcut shape for reproducing the ancient book of *hanpon*. First, using commercial software developed from the algorithm[4], we measure 3D of the woodcut, whose data are to be aligned to make the 3D digital woodcut. After the woodcut shape is restored, we can print out virtual *hanpon*.

2. Measuring and aligning 3D Woodcut point cloud data

The non-contact 3D-Measurement machine VIVID910 is utilized to measure woodcut point clouds. When measuring the woodcut, it is necessary to take bump patterns on the woodcut surface as precisely as possible. Because the measurement machine cannot measure all the surface of the object at once, we need to generate object shape models by using measured point clouds from multiple viewpoints. Commercial software is utilized to align cloud data. At first, the rough corresponding point is defined on different cloud data. Then, the software can value the globe errors of the total aligned point clouds and adjust the position and direction of these point clouds. At the end, the point clouds are aligned in one 3D digital woodcut model. Figure 2 show the aligned result. Different colors show different parts of the woodcut cloud point data.

3. Constructing wood board model

As the woodcut surface is very delicate, the computing cost will be considerable if one uses this delicate model. A rough mesh model is constructed to restore the woodcut shape. As the woodcut is made from a piece of wood board, if the wood board shape is restored, so will be the woodcut shape. Hence, the rough model is the wood board. This wood board model will is a box after its shape is restored.

To construct the wood board model, we utilize the woodcut section. In the Figure 3, the bottom line is the woodcut model section, and the top line is the constructed wood board model section. The top line is the connection line between local highest points of woodcut model sections. Some woodcut sections are utilized to construct the total wood board model. The distance between these sections is the same as the woodcut's thickness. After all sections are processed with this method, we can obtain the wood board for restoring.

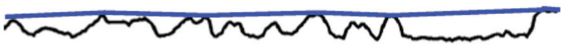


Fig. 3 Constructing the wood board model from Woodcut section

4. Restoring the wood board model

As mentioned above, wood is an orthotropic material. The wood study needs to consider the three main axes of the wood. Figure 4 shows a tree trunk form in a coordinate system with the three standard directions: the trunk axis or fiber direction *L*, the radial direction *R* that passes through the tree's core, and the tangent direction *T* along the annual ring. When wood dries and shrinks, shrinkage in the tangential, radial, and fiber directions occurs in a ratio of 10:5:0.5. With this reason, the wood board distorts.

While the surface of woodcut is black, on the woodcut end, we can find some exposed parts, including the annual ring pattern. From this annual ring, the tree's core can be understood as the annual ring is nearly a concentric cylinder. From the position of the tree's core, the wood three standard directions can be determined all over the wood board.

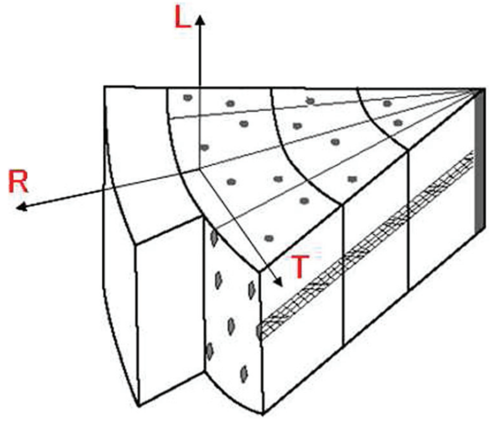


Fig. 4 Wood three axis.

The woodcut is made of wild cherry tree or boxwood. Shrinkage of this type of the wood is about 0.31% on the tangent direction *T* , and 0.17% on the radial direction *R*. Since the shrinkage on fiber direction *L* is very small, its change *L* is ignorable. If the moisture content is lower than 30%, the wood starts distorting, in Japan, after the moisture content is 15%, the moisture content stops changing, and the shape stops distorting. As Figure 5 shows, point *A* is one point in the wood board. *ε* is the vector to show plastic strain on the point *A*. *T* is tangent direction and *R* is radial direction. The plastic strain can be shown as the following equation:

$$\varepsilon = \varepsilon_T - \varepsilon_R =(S_T \times T - S_R \quad R) \quad \Delta W \, (1)$$

where, *ST* is shrinkage on direction *T* and S_R is shrinkage on direction *R*. Δ*W* is moisture content variation and is 1%. Using this equation, the plastic strain ε can be computed. The restoring process is repeated until the wood board surface gets closer to a plane or the total moisture content reaches 30%. This is the restored wood board.

As the connection between the woodcut point cloud and wood board mesh is known, it is easy to obtain a restored woodcut point cloud from the restored wood board. As the local highest point is on a plane after shape restored, the virtual *hanpon* can be printed easily by using the height information of the restored woodcut. If the height is higher than 0.3mm under the highest point, it is the part for printing. The printed result is shown in Figure 6.

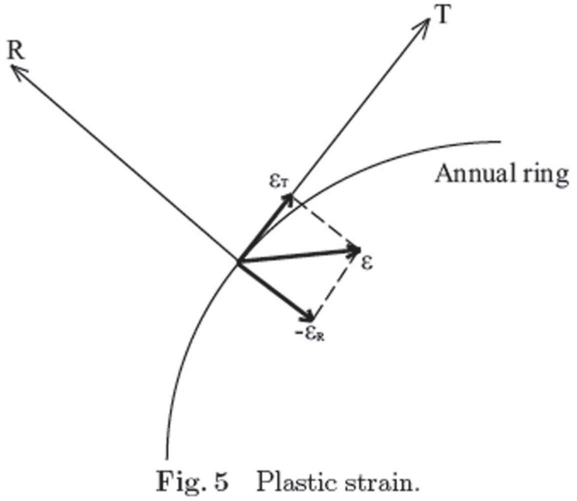


Fig. 5 Plastic Strain.

5. Results

The final rendering result is carried out on a computer with the GPU (Graphics Processing Unit) and can render the *hanpon* on real time. The graph card is NVIDIA GeForce 6800 GS.

Since the Japanese paper became old and was not as white as it was printed out, brown color captured from old Japanese paper is added into the virtual printing result. The virtual printing result and the Japanese fiber model are utilized to render based on the fiber reflection model.[5] Figure 6 shows the final rendering results. The left image shows the *hanpon* appearance with white color after it was printed out about 200 years ago. The right one shows the result after *hanpon* color variation. Using the technique proposed, the ancient *hanpon* is represented in the virtual world.

As the woodcut shape is restored, the size of printed *hanpon* is the same as the original *hanpon* when it was printed hundreds years ago. Comparing the size of *hanpon* and virtual printed result using proposed method, the published history of *hanpon* can be understood well. This information is very important for history research.

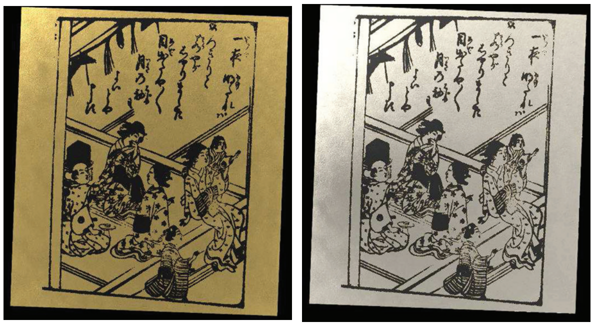


Fig. 6 Rendering result of hanpon.

6. Conclusion and discussion

This paper proposes the techniques to restore woodcut shapes and reproduce ancient *hanpon*. The bump pattern of woodcut surface is very minute, and the color of woodcut is black. To restore the woodcut shape, therefore, the FEM method is utilized. This method can obtain fine printing results even woodcut is distorted. To our knowledge, it is first time to try to study *hanpon* history using virtual Woodcut printing method.

Some work, however, needs to be done in the future. Right now, using 3D scanner cannot capture the woodcutfs details. Thus, we need to develop techniques to improve original 3D data precision. One idea is to use the high resolution camera to take high resolution photos in a different lighting environment. Using image processing technique makes the printed areas possible to be extracted. Then, using the proposed techniques in this paper to calibrate the *hanpon* size, we can obtain delicate printed results. We also hope to develop a Virtual Reality system in which the user can watch the cultural heritage and touch its surface at the same time. This Virtual Reality system is not only a new type digital museum, but also is entrainment system such as game as well.

References

1) M.Okada, S.Mizuno and J.Toriwaki: Virtual Sculpting and Virtual Wood block Printing by Model-Driven Scheme, *the Journal of the Society for Art and Science*, Vol.1, No.2, pp.74–84 (2002).

2) Yin, X., Eto, T., Akama, R., Nagai, K. and Tanaka, H.T.: Digital Woodcut Measurement and Ancient Hanpon Rendering, *Proceedings of 2008 ASIA GRAPH*, Vol.2, No.1, pp.31–36 (2008).

3) Carlsson, P. and Esping, B.: Optimization of the wood drying process, *Structural and Multidisciplinary Optimization*, Vol.14, No.4, pp.232–241 (1997).

4) Soucy, M. and Laurendeau, D.: A General Surface Approach to the Integration of a Set of Range Views, *IEEE Trans. Pattern Anal*, Vol.17, No.4, pp.344–358 (1995).

5) Yin, X., Cai, K., Takeda, Y., Akama, R. and T.Tanaka, H.: Measurement of Reflection Properties in Ancient Japanese Drawing Ukiyo-e, *Proceedings of 8th Asian Conference on Computer Vision (ACCV 2007), Part I, LNCS 4843*, pp.779–788 (2007).

Author Index