The Alliance of Digital Humanities Organisations
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 interactif
DIGITAL HUMANITIES 2011
CONFERENCE ABSTRACTS
DH
2011
STANFORD, CA

The Alliance of Digital Humanities Organizations
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 interactif

Digital Humanities 2011

Conference Abstracts

Stanford University, Stanford, CA, USA
June 19 – 22, 2011


DH
2011

The 23rd Joint International Conference of the Association for Literary and Linguistic Computing and Association for Computers and the Humanities

and

The 4th Joint International Conference of the Association for Literary and Linguistic Computing, the Association for Computers and the Humanities and the Society for Digital Humanities – Société pour l'étude des médias interactif

International Program Committee

- Arianna Ciula (ALLC)
- Dominic Forest (SDI-SEMI)
- Cara Leitch (SDI-SEMI)
- John Nerbonne (ALLC)
- Bethany Nowviskie (ACH)
- Daniel O'Donnell (SDI-SEMI)
- Dot Porter (ACH)
- Jan Rybicki (ALLC)
- John Walsh (ACH)
- Katherine Walter (ACH: Chair)

Local Organizing Committee

- Suzanne Bennett, Conference Services
- Philip Gin, Conference Services
- Ryan Heuser, Volunteer Coordinator
- Matthew Jockers, Local Host
- Tanya Walker, Conference Services
- Melanie Walton, Conference Services
- Glen Worthey, Local Host

Conference Sponsors

- Alexander Street Press
- Ashgate Publishing Group
- The Bill Lane Center for the American West at Stanford University
- Gale Publishing
- HathiTrust Research Center
- ProQuest
- The Stanford University Department of English
- Stanford University Libraries and Academic Information Resources
- The Stanford Literary Lab

Conference Volunteers

- Benjamin Albritton
- Nicole Coleman
- Kimberley Hayworth
- Long Le-Khac
- Elijah Meeks
- Kathryn VanArendonk

ISBN 978-0-911221-47-3
Published by Stanford University Library
Conference logo and cover design by Nicole Coleman

A Welcome from the Stanford University Librarian

Michael Keller
Stanford University

Stanford University
LIBRARIES & ACADEMIC INFORMATION RESOURCES

Digital Humanists meeting at Stanford this month, whether present virtually or physically, are most welcome to our campus. Your host organization this year is Stanford University Libraries and Academic Information Resources, a long name for a complex and unique organization in higher education. SULAIR is a library, a cybrary, an academic computing organization, a publisher of scholarly monographs (Stanford University Press), and a provider of publishing services to about 140 scholarly journal publishers (HighWire Press), each specialty division supporting and informing the others in this mélange.

We have supported digital humanities research and teaching numerous ways over the years through our digitization services, our academic technology specialist agents, our humanities curators, our digital archiving programs (LOCKSS and the Stanford Digital Repository), and our e-publishing services. We have found collaborators and fellow travelers along this route from the British Library, the Matthew Parker Library of Corpus Christi College of the University of Cambridge, the Bibliothèque nationale de France, the Bibliotheca Alexandrina, University of Virginia, University of Michigan, the University of Illinois at Urbana-Champaign, and the University of Alicante, among others.

A fluff piece in American Libraries last year described SULAIR as a “juggernaut of innovation.” Perhaps we are, but our collective nose has been on the grindstone employing Blacklight to provide a richer OPAC, including virtual browsing of our classified collections and partnering with Google not just in the library digitization program, but as well in prototyping an innovative e-Thesis service integrating academic decisions on dissertations, easy workflows leading digital dissertations to our digital archive (the Stanford Digital Repository) and ultimately to Google for free indexing and access on a global basis. Matt Jockers and others in the SEASR team will report on the NLP research underway here and with several collaborating institutions. This recital could go on and on.

We here use i.t. for prosaic purposes, getting unusual returns on investments in clever uses of technology in support of the usual library/cybrary functions. We use i.t. for innovative purposes, often resulting in open source tools and environments (Hydra and LOCKSS, for example). And we promote and support the use of i.t. in research and teaching (digital maps & GIS in numerous projects, and CourseWork, the local Sakai implementation, for example).

You are all most welcome here. Glen Worthey, chief of our 25-year-old Humanities Digital Information Service, and Matt Jockers, Academic Technology Specialist and leader of the SEASR project, are responsible for local arrangements for this meeting. We are excited as they are to have you here and look forward to learning from all of you.

Mike

Michael A. Keller
University Librarian
Director of Academic Information Resources
Founder/Publisher HighWire Press
Publisher Stanford University Press

Welcome to the Big Tent

Matthew Jockers
mjockers@stanford.edu
Stanford University

Glen Worthey
gworthey@stanford.edu
Stanford University

Welcome to Stanford, and welcome to Digital Humanities 2011. Our very public goal over the past two years (and more) of preparation for this week has been to bring this, our favorite DH conference, to Stanford. But we also confess to another, more private goal: to bring Stanford to the DH conference. We've long recognized the digital humanities practices, and even a very particular digital humanities spirit, in the work of many, many of our Stanford colleagues; we've long suspected that they would find a welcome place in the community represented by ADHO and the annual DH conferences — even though, until recently, only a handful of us have been regular DH attendees.

Every institution has its quirks, its culture and its subcultures, and one aspect of Stanford academic life that we've grown accustomed to here is what we've come to call the "entrepreneurial," by which of course we mean enterprising, risk-taking, and adventuresome. (Of course, it's also a nod to Stanford history and geography, and to the many generations of Stanford's Silicon Valley brainchildren, from pre-Hewlett-Packard to post-Google.) But "entrepreneurial" means not only that: Stanford's digital humanists have not until recently been much engaged in the formation of DH centers or departments, nor have they coalesced around a single professional society or journal or annual conference. Instead, our work and our practitioners have been distributed across many academic departments, in the Libraries, and various research initiatives, with pockets of intense and important DH activity in all those places.

The theme we've chosen for DH2011 is "Big Tent Digital Humanities," and this was meant, in part, to convey our own desire to include in the DH2011 the many different varieties of DH practice that surround us right here at Stanford. At the same time, the "big tent" seemed to us an appropriate metaphorical response to some of the debates that have flourished in the digital humanities worldwide, especially in the past few years, about the meanings and limits of the Digital Humanities designation. Although DH2011 belongs to everyone who participates in it, by choosing this theme we meant to announce publicly our own opinion that a broad and diverse and vibrant DH field, one in which a thousand flowers might bloom, is the sort of DH we at Stanford believe in and hope to promote.

(And speaking of a thousand flowers: the psychedelic flower-child theme of our conference website, our logo, and the design of the book of abstracts you're reading now, were all created by our many-talented Stanford DH colleague Nicole Coleman. We hope it conveys to you not only our heritage of California dreaming, but also — and more importantly — our sense of wonder and appreciation for the many-splendored field of DH, for its practice of creative exuberance, for its opening of the scholarly mind and senses to new and revolutionary ways of seeing and thinking about the humanities. We believe you'll find that the DH revolution we hope to promote is of a peace-loving, sunshiny nature — but no less revolutionary for all that.)

A proper list of acknowledgments would run to many, many pages, but let us mention a particular few without whose help both we and DH2011 would be utterly lost: the ADHO Conference Coordinating Committee chair, John Unsworth, has helped us through every stage of the planning process, from well before the submission of our bid to host DH2011 more than two years ago, to.... Well, I imagine we'll have still more questions for John and other ADHO executives long after everyone else has gone home. ADHO's International Program Committee, chaired with diligence and a gentle hand by Kay Walter, has assembled a wonderful bouquet of papers, panels, posters, and workshops from a select portion of the many hundreds of DH flowers that bloomed in the submission process; of course, Kay and her team also had the thankless task of having to turn away many worthy proposals. To all those

who submitted, we express our sincere thanks; to those many whose proposals did not end up in the final program, we also say: We know how you feel; we've been there too. Even the biggest tents have sell-out crowds.

Closer to home, a small army of people have helped make this conference possible, foremost among them our meeting planner Melanie Walton, with the help of a great team in Stanford Conference Services. Melanie and company have expertly handled more details of more aspects of pulling off a conference like this than the two of us even knew existed, and they've done so with patience and good humor. Our team of conference volunteers is made up of good-humored graduate students from the Stanford Literary Lab, under the direction of Franco Moretti and Matt (one of yours truly), as well as a number of colleagues from the Stanford University Libraries. These people have helped with everything from text markup to staffing the registration tables, but they're really far too accomplished to be doing this sort of thing; ask them about their *real* work as they're showing you around campus or helping you with WiFi.

Of course, a lot more could be said about Stanford, about DH in general, and about DH2011 in particular (not to mention at least a dozen more clichés of 1960s counterculture, and a dozen different ways of parsing our "big tent" metaphor) — but let's get on with the show.

Matt Jockers & Glen Worthey, your local hosts

From the Chair of the DH2011 International Program Committee

Katherine L. Walter
kwalter@unlnotes.unl.edu
University of Nebraska-Lincoln

This year is a very special one as the Digital Humanities 2011 conference returns to California with the new theme "Big Tent Digital Humanities." The DH2011 conference has more papers and a wider range of topics than many in the past. In large part, this is thanks to all of you who took the theme to heart and participated by submitting abstracts for panels, papers and posters. In hopes of addressing the issue of a growing conference, the program committee called for both long and short papers, with longer ones addressing newly concluded research or theory, and short papers reporting on research in progress. It is through the short papers that we were able to expand the conference. The Program Committee encourages you to explore new topics outside your comfort zone. Our field is broad and the conference reflects this. Where better to learn new things?

As mentioned, the Digital Humanities conference is growing, and the number of papers proposed far exceeded the number of time slots available. In response to this, the Alliance of Digital Humanities Organizations, its member associations, and the Conference Coordinating Committee are considering new approaches to the conference. Among ideas raised have been expanding the number of concurrent sessions, extending the conference by one day, and extending the length of the day. Your thoughts on this issue are appreciated and can be submitted to the Conference Coordinating Committee.

In recognition of our Big DH Tent, our keynote addresses are by David Rumsey, creator of the well-known eponymous map collection and president of Cartography Associates; and J.B. Michel and Erez Lieberman-Aiden of Harvard University, who will be speaking about their large-scale "culturomics" research conducted in association with the Google Books project. Thus the conference begins with the public (and geospatial) humanities, and concludes with corpus research on previously unimaginable scales.

I would like to thank the members of the international program committee who have contributed so much of their time. The members of the committee are: Arianna Ciula (ALLC) Dominic Forest (SDH-SEMI) Cara Leitch (SDH-SEMI) John Nerbonne (ALLC) Bethany Nowviskie (ACH) Daniel O'Donnell (SDH-SEMI) Dorothy (Dot) Porter (ACH) Jan Rybicki (ALLC) John Walsh (ACH)

Local organizers Glen Worthey and Matt Jockers have been great contributors to our efforts.

Finally, I would like to thank all of you who agreed to review abstracts; you are acknowledged individually in the pages of this book. Your time is appreciated, and together, I believe we have planned an exceptional conference.

Charles Douglas Bush, 1948 - 2011

The digital humanities community lost a good friend and colleague with the death of Chuck Bush, long-time Treasurer of the Association for Computers and Humanities and an unfailingly warm presence at the annual digital humanities conference which he attended with his wife, Junola. Chuck was trained as a linguist and worked at Brigham Young University as a Senior Research Consultant in the Humanities Technology and Research Support Center, where he often explored his interests in both print and electronic publishing. He was one of the early advocates for Humanities Computing education, as he both administered and taught in the Computers and the Humanities Program. Humanities faculty at BYU will remember Chuck for his kind and dedicated assistance with arcane computer issues, as well as for his enthusiasm and unfailingly pleasant manner.

Chuck served for more than 20 years on the ACH Executive Council, and for those of us who worked closely with him during that time he was a constant source of calm good sense, humor, and organizational memory. An absolutely trustworthy steward for the ACH during its slowly rising fortunes, Chuck played a crucial role in managing the ACH's alliance with ADHO in 2005-6 and in helping the organization through a challenging transitional time. He could be counted on to suggest a helpful compromise, clarify a tricky point, or bring a discussion down to earth-he was never rattled or partisan.

Chuck was not given to self-display, but he had a great and subtle wit even in routine email. He was principled without imposing on others. With his particular dedication to supporting younger scholars and education, he was a strong champion of the ACH and ADHO bursary programs, and as ACH Treasurer he had the pleasure each year of handing out checks in a worthy cause. Those of us who met him annually for meetings and a conference will remember him with great affection, taking off his glasses in a meditative way before offering his thoughts. He will be much missed and warmly remembered by all.

Prepared by Julia Flanders with contributions from the DH community.

Table of Contents

Papers

Posters

List of Reviewers

- Akama, Hiroyuki
- Anderson, Deborah
- Anderson, Jean Gilmour
- Anderson, Sheila
- Andreev, Vadim Sergeevich
- Baayen, Rolf Harald
- Barney, Brett
- Battino Viterbo, Paolo
- Bauman, Syd
- Baumann, Ryan Frederick
- Bearman, David
- Beavan, David
- Bellamy, Craig
- Bennis, Hans
- Bentkowska-Kafel, Anna
- Bia, Alejandro
- Biber, Hanno
- Blanke, Tobias
- Bodard, Gabriel
- Boggs, Jeremy
- Bolter, Jay David
- Booij, Geert E.
- Borin, Lars
- Bosse, Arno
- Boves, Lou
- Bowen, William
- Bradley, John
- Brey, Gerhard
- Brown, Susan
- Burnard, Lou
- Burr, Elisabeth
- Bush, Chuck
- Cantara, Linda
- Carson, Christie
- Caton, Paul
- Cayless, Hugh
- Chen, Shih-Pei
- Chesley, Paula Horwath
- Ciula, Arianna
- Clement, Tanya
- Conner, Patrick
- Connors, Louisa
- Cooney, Charles M.
- Cooper, David Christopher
- Cossard, Patricia Kosco
- Craig, Hugh
- Cummings, James C.
- Cunningham, Richard
- Dahlstrom, Mats
- David, Stefano
- Dawson, John
- Devlin, Kate
- Dik, Helma
- DiNunzio, Joseph
- Dombrowski, Quinn Anya
- Downie, J. Stephen
- Dubin, David S.
- Dunn, Stuart
- Durand, David G.
- Durusau, Patrick
- Eberle-Sinatra, Michael
- Eder, Maciej
- Edmond, Jennifer C
- Egan, Gabriel
- Eide, Øyvind
- Ell, Paul S
- Esteva, Maria
- Everaert, Martin
- Fiormonte, Domenico
- Fischer, Franz
- Fitzpatrick, Kathleen
- Flanders, Julia
- Flatscher, Markus
- Forest, Dominic

- Fraistat, Neil R.
- France, Fenella Grace
- French, Amanda
- Fritze, Christiane
- Funkhouser, Chris
- Furuta, Richard
- Galina Russell, Isabel
- Gallet-Blanchard, Liliane
- Galloway, Patricia
- Gants, David
- Gärtner, Kurt
- Gartner, Richard
- Gilbert, Joseph
- Giordano, Richard
- Goldfield, Joel
- Gow, Ann
- Grob, Nathalie
- Gueguen, Gretchen Mary
- Hanlon, Ann
- Hanrahan, Michael
- Harbeson, Eric
- Harris, Katherine D.
- Hawkins, Kevin Scott
- Heiden, Serge
- Hernández Figueroa, Zenón
- Hirsch, Brett
- Hockey, Susan
- Holmes, Martin
- Hoover, David L.
- Hswe, Patricia
- Hughes, Lorna
- Huitfeldt, Claus
- Hulk, Aafke
- Hunyadi, László
- Hyman, Malcolm D.
- Isaksen, Leif
- Ivanovs, Aleksandrs
- Jockers, Matthew
- Johnsen, Lars
- Johnson, Ian R.
- Juola, Patrick
- Kaislaniemi, Samuli
- Kansa, Eric Christopher
- Kansa, Sarah Whitcher
- Keating, John Gerard
- Kelleher, Margaret
- Khosmood, Foaad
- Kirschenbaum, Matthew
- Kraus, Kari michaele
- Krauwer, Steven
- Kretzschmar, William
- Krot, Michael Adam
- Lancaster, Lewis Rosser
- Lavagnino, John
- Lavrentiev, Alexei
- Leitch, Caroline
- Lendvai, Piroska
- Lewis, Benjamin G.
- Litta Modignani Picozzi, Eleonora
- Llewellyn, Clare
- Lombardini, Dianella
- Lüngen, Harald
- Luyckx, Kim
- Mahony, Simon
- Makinen, Martti
- Mari, Francesca
- Martin, Worthy N.
- Martinet, Marie-Madeleine
- McPherson, Tara
- Mealand, David
- Meister, Jan Christoph
- MendezRodriquez, Eva
- Meschini, Federico
- Miles, Adrian
- Miyake, Maki
- Mostern, Ruth

- Moulthrop, Stuart
- Mylonas, Elli
- Myojo, Kiyoko
- Nagasaki, Kiyonori
- Nelson, Brent
- Nerbonne, John
- Neuman, Michael
- Newton, Greg T
- Nieves, Angel David
- Norrish, Jamie
- Nowviskie, Bethany
- Nyhan, Julianne
- O'Donnell, Daniel Paul
- Olsen, Mark
- Opas-Hänninen, Lisa Lena
- Ore, Christian-Emil
- Ore, Espen S.
- Pantou-Kikkou, Eleni
- Parker, Alexander
- Pasanek, Brad
- Pierazzo, Elena
- Piez, Wendell
- Pitti, Daniel
- Porter, Dorothy Carr
- Priani, Ernesto
- Pytlik Zillig, Brian L.
- Rahtz, Sebastian
- Rains, Michael John
- Ramsay, Stephen
- Rehbein, Malte
- Rehm, Georg
- Renear, Allen H.
- Reside, Doug
- Robertson, Bruce
- Robey, David
- Robinson, Peter
- Rockwell, Geoffrey
- Rodríguez, Nuria
- Roe, Glenn H
- Romary, Laurent
- Roueché, Charlotte
- Roued-Cunliffe, Henriette
- Rudman, Joseph
- Ruecker, Stan
- Russo, Angelina
- Rybicki, Jan
- Saint-Dizier, Patrick
- Sánchez Quero, Manuel
- Sanz, Concha
- Scheinfeldt, Joseph Thomas
- Schlitz, Stephanie
- Schmidt, Harry
- Schmidt, Sara A.
- Schreibman, Susan
- Seppänen, Tapio
- Shaw, William Stewart
- Siemens, Lynne
- Siemens, Raymond George
- Simons, Gary F.
- Sinclair, Stéfan
- Singer, Kate
- Smith, David A.
- Smith, Natalia (Natasha)
- Snyder, Lisa M.
- Spence, Paul Joseph
- Sperberg-McQueen, Michael
- Spiro, Lisa
- Sternfeld, Joshua
- Stokes, Peter Anthony
- Sukovic, Suzana
- Suzuki, Takafumi
- Swanstrom, Elizabeth Anne
- Tabata, Tomoji
- Thaller, Manfred
- Tripp, Mary L.
- Tufis, Dan

- Unsworth, John
- Uszkalo, Kirsten Carol
- Van den Branden, Ron
- van den Herik, H. J.
- Van Elsacker, Bert
- Váradi, Tamás
- Walker, Brian David
- Walsh, John
- Walter, Katherine L.
- Warwick, Claire
- Wendrich, Willeke
- Wilkens, Matthew
- Willett, Perry
- Winder, William
- Witt, Andreas
- Wittern, Christian
- Wolff, Mark
- Worthey, Glen
- Yu, Bei
- Zafrin, Vika
- Zimmerman, Matthew

Plenary Sessions

Re-Imagining Scholarship in the Digital Age

Gaffield, Chad

1. Abstract

Quite unexpectedly and often in surprising ways, Digital Humanities have been playing a central role in the larger re-imagining of scholarship in the early 21st century. This re-imagining is transforming teaching, research and, indeed, all aspects of academic life. Moreover, the established boundaries between, and relationships among, scholarly activities on campus and in the larger society are falling flat or being re-configured through networks, clusters and dynamic forms of engagement. The result is exhilarating and un-nerving, inspiring and challenging, energizing and exhausting, if judged by public debate on campus and beyond about the changing post-secondary landscape. But if we focus on the emergence and current trajectory of Digital Humanities, we can perceive with cautious optimism the ways in which a re-imagined scholarship is beginning to enhance learning, to help interpret the past and present, and to contribute to meaningful life in the 21st century.

2. Bio

Gaffield is Professor of History at the University of Ottawa and currently on leave while he serves as President of the Social Science and Humanities Research Council of Canada. Founding Director of the Institute of Canadian Studies, Gaffield has been since the 1970s at the forefront of computer-based analyses of long-term social change. He has played a leading role in, and produced award-winning publications from, database projects such as the Canadian Social History Project, the Vancouver Island Project, the Lower Manhattan Project, and the Canadian Families Project; as President of the Humanities and Social Sciences Federation of Canada, he also championed the Data Liberation Initiative.

Among his many notable accomplishments, the prize is awarded to Gaffield for his role as Principal Investigator for the Canadian Century Research Infrastructure project (CCRI; www.ccri.uottawa.ca). CCRI has created a foundation for the study of social, economic, cultural, and political change at a national level, beginning with digital reconstruction of censuses that sit at the core of a pan-national research database consisting of pertinent contextual data drawn from newspapers, parliamentary proceedings, legislative records and beyond.

Culturomics: Quantitative Analysis of Culture Using Millions of Digitized Books

Lieberman-Aiden, Erez

Michel, Jean-Baptiste

1. Abstract

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of ‘culturomics,’ focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

2. Bios

Erez Lieberman Aiden is a fellow at the Harvard Society of Fellows and Visiting Faculty at Google. His research spans many disciplines and has won numerous awards, including recognition for one of the top 20 "Biotech Breakthroughs that will Change Medicine", by Popular Mechanics; the Lemelson-MIT prize for the best student inventor at MIT; the American Physical Society's Award for the Best Doctoral Dissertation in Biological Physics; and membership in Technology Review's 2009 TR35, recognizing the top 35 innovators under 35. His last three papers - two with JB Michel - have all appeared on the cover of Nature and Science.

Jean-Baptiste Michel is FQEB Fellow at Harvard and Visiting Faculty at Google. With Erez Lieberman Aiden, he founded the Cultural Observatory at Harvard, where their team develops quantitative approaches to the humanities and social sciences. Jean-Baptiste is an Engineer of Ecole Polytechnique, and received an MS in Applied Math and a PhD in Systems Biology from Harvard.

Reading Historical Maps Digitally: How Spatial Technologies Can Enable Close, Distant and Dynamic Interpretations

Rumsey, David

1. Abstract

Maps are dense, complex information systems arranged spatially. While they share similarities with other visual artifacts, their uniqueness as spatially arranged visual information both allows for and demands special digital approaches to understand and reuse their content. Georeferencing, vectorization, virtual reality, image databases, and GIS-related tools all work to unite our eyes, minds, and computers in new ways that can make historical maps more valuable and accessible to humanists concerned with place and space over time. Rumsey will explore the tools and techniques that have implications for the ways digital humanists approach visual information.

2. Bio

David Rumsey is a renowned collector of historical maps, a digital librarian, an online publisher, builder, and philanthropist. His collection of more than 150,000 maps is one of the largest private map collections in the United States, and he recently announced his intention to donate it to the Stanford University Libraries. With his growing online collection of more than 26,000 maps, available to all in high resolution and with expert cataloging, Rumsey is one of the most visible and important modern distributors of historical treasures for the common good, a pioneer Internet philanthropist, and a public Internet intellectual. Visit the David Rumsey Map Collection online at http://www.davidrumsey.com/. With his bold experiments in the use of GIS with historical maps, his innovative use of virtual worlds for purveyance of serious scholarly materials, and his outspoken and concrete actions toward the building of a real public digital library, David Rumsey is a rare and exemplary figure of antiquarian in the digital world, and entrepreneur in the academy.

Pre-conference Workshops

Visualization for Literary History

Brown, Susan
Susan.Brown@ualberta.ca
English and Humanities Computing, University of Alberta/University of Guelph

Ruecker, Stan
sruecker@ualberta.ca
English and Humanities Computing, University of Alberta

Rockwell, Geoffrey
grockwel@ualberta.ca
Philosophy and Humanities Computing, University of Alberta

Sinclair, Stéfan
sgs@mcmaster.ca
Communications and Multimedia, McMaster University

This workshop will present, demonstrate, and provide participants with the opportunity to test and discuss prototypes of several experimental visualization tools for literary studies. The tools will provide a range of approaches to visualizing the Orlando Project's textbase. Some but not all will allow for input of other data. Although the workshop is focused on literary studies in English, we welcome participants from other related disciplines such as history, philosophy, the history of science, media studies, or library and information science, as well as those with an interest in text visualization generally, and those interested in corpuses in languages other than English.

The Orlando Project's fifteen-year experiment in literary history explores the potential of computers to support new modes of humanities research, particularly the potential of digital technologies to enable interpretive and critical scholarship. The major result of that endeavour, the online *Orlando: Women's Writing in the British Isles from the Beginnings to the Present* (Brown et al 2006; orlando.cambridge.org), constitutes the single most extensive and detailed resource in the area, hailed by the Modern Language Association's *Guide to Literary Research* as "a model for similar databases that will supplant printed literary dictionaries, encyclopedias, and handbooks" (Harner). Though *Orlando* resembles a reference work, its electronic structure embeds an entire critical and theoretical framework to support advanced literary historical enquiry. The workshop proposed here will present and allow participants to experiment with prototypes based on emergent methods in text mining and visualization that leverage that embedded structure to enable new discovery paths in literary history.

The *Orlando* textbase—about 80 print volumes' worth of born-digital scholarship encoded with an XML tagset of more than 250 tags covering the production, characteristics, and reception of texts)—constitutes a rare testbed for investigating the mining of structured text. Its online interface and search system were developed according to W3C standards to exploit the underlying markup, and designed to meet the expectation of text-oriented users of conventional online tools. This existing interface is very search-oriented and entirely textual in its delivery of results.

Current research in humanities computing and human-computer interaction is increasingly expanding beyond the text-oriented information retrieval paradigm, to explore instead the many opportunities offered by new, more flexible, more visually-oriented platforms for web delivery (e.g. Ahlberg and Shneiderman 1994; Bederson 2000, 2001; Harris 2006, 2007; Greengrass and Hughes, 2008). In this period of transformation, the scholarly interface requires not only experimentation but also careful assessment to see what works to make digital materials of real value to humanities scholars. As argued by Ramsay (2003), Unsworth (2006), and others, using computers to do literary research can contribute to hermeneutic or interpretive inquiry. Digital humanities research has inherited from computational science a leaning towards systematic knowledge representation. This has proved serviceable in some humanities activities, such as editing, but digital methods have far more to offer the humanities than this. As Drucker and Nowviskie have argued, "The computational processes that serve speculative inquiry must be dynamic and constitutive in their operation, not merely procedural and mechanistic" (431).

Our goal for this workshop is to provide those interested in literary studies and the digital humanities with an introduction to some of the tools being developed to support interactive speculative inquiry through text mining and visualization. In the process, we hope to garner insight into users' reactions to these tools to inform further design and development activities. The prototypes presented at this workshop are being developed as possible interfaces to complement Orlando's current, more conventional one.

The prototypes presented at the workshop will include the following:

- Mandala Browser: this browser allows users to create “magnets” based on free text or XML search that attract to them items in a text collection, and to visualize the relationships between different sets. It can be used with the Orlando data or with other textual datasets. (Sinclair and Ruecker)
- Orlando Degrees of Separation tool: this tool shows the connections between individuals in the Orlando data by way of other people, places, organizations, or titles. The challenge is in organizing the visualization of the paths when there are multiple ones, as there frequently are in this highly interlinked set of data
- OrlandoVision, a network graph visualization tool: creates a social network graph in which individuals’ names are nodes and links between them are edges, which are color-coded according to the semantic context of the link as represented in the markup
- Breadboard interface for tracing links between individuals and entities: a more textually-oriented interface for browsing links between individuals and entities within the Orlando data
- Voyeur: a general-purpose web-based text analysis environment designed for large-scale corpora; includes experimental visualization modules for exploring word trends, named entities, and other textual features
- this tools shows the connections between individuals in the Orlando data by way of other people, places, organizations, or titles. The challenge is in organizing the visualization of the paths when there are multiple ones, as there frequently are in this highly interlinked set of data
- possibly other visualization tools emergent from current research: we are experimenting with other mining and visualization tools between now and DH2011 and may pull ones that seem to have potential into the workshop program

This workshop emerges from ongoing research on visualization for literary research, and participants will be asked, but not required, to participate in the study through surveys, interviews, and recording of user sessions in accordance with the ethics protocols approved by our respective universities.

References

Ahlberg, C., Shneiderman, B. (1994). 'The Alphaslider: A compact and rapid selector'. Conference proceedings on human factors in computing systems: “celebrating interdependence". .

Bederson, B. (2001). *PhotoMesa: A zoomable image browser using quantum treemaps and bubblemaps. Proceedings of the 14th annual ACM symposium on user interface software and technology.* . http://doi.acm.org/10.1145/502348.502359.

Brown, Susan (2006a). *Orlando: Women's Writing in the British Isles from the Beginnings to the Present.* Clements, Patricia, Grundy, Isobel (eds.). . http://orlando.cambridge.org.

Drucker, J., Nowviskie, B. (2004). *Speculative computing: Aesthetic provocations in humanities computing. A Companion to Digital Humanities.* Schreibman, S., Siemens, Ray, Unsworth, John (eds.). .

Greengrass, Mark, Hughes, Lorna (eds.) (December 2008). *The Virtual Representation of the Past.* .

Harner, James L. (2008). *Literary Research Guide: An Annotated Listing of Reference Sources in English Literary Studies, 5th edition.* .

Harris, J. (2006). '10 by 10: 100 words and pictures that define the time'. http://www.tenbyten.org/10x10.html.

Ramsay, Stephen (2003). 'Toward an Algorithmic Criticism'. *Literary and Linguistic Computing.* .

Sinclair, Stéfan, Ruecker, Stan (2008). 'Mandala Rich Prospect Browser'. http://mandala.humviz.org.

Unsworth, John (2006). 'New methods for Hmanities Research'. http://www3.isrl.uiuc.edu/~unsworth/lyman.htm.

Introductory TEI ODD

Cummings, James
james.cummings@oucs.ox.ac.uk
Oxford University Computing Services

Rahtz, Sebastian
sebastian.rahtz@oucs.ox.ac.uk
Oxford University Computing Services

---

1. Abstract

ODD (One Document Does it all) is the XML vocabulary which the Text Encoding Initiative has developed to describe itself, and which users of the TEI employ to customize the TEI and create documentation and schemas appropriate to their varied needs. It can be used to document and describe any XML vocabulary. Attendees are expected to have some basic knowledge of XML and the TEI.

Introductory TEI ODD is a three-hour tutorial taught by the TEI@Oxford team (James Cummings and Sebastian Rahtz) on using TEI ODD for documenting and constraining your project's TEI XML. This beginners-level course provides a hands-on practical introduction to the basic ideas of the TEI ODD system, exploring the process of designing a TEI profile from the user perspective, and using the TEI's Roma web application to model a schema. We will explore the full capabilities of Roma, from designing and testing a basic XML schema with a few clicks, up to the design of a highly customized and multilingual application profile.

Participants will be taught using Roma and the oXygen XML editor. They will define a detailed customization of the TEI, generate schemas and documentation, and have an understanding of the TEI ODD language. This tutorial will equip participants with the necessary skills to customize the TEI for their project's specific needs.

2. Tutorial Structure

1. Talk 1: An introduction to TEI ODD concepts and the Roma Web Application (30min)
2. Exercise 1: Creating your first TEI customization (30 min)
3. Talk 2: More complicated TEI customizations (30 min)
4. Exercise 2: Constraining and Extending the TEI for your Project (30 min)
5. Talk 3: The TEI ODD Format: What is happening underneath? (30 min)
6. Exercise 3: Editing the underlying TEI ODD files (30 min)

3. Contact Information

Dr James Cummings, Oxford University Computing Services, 13 Banbury Road, Oxford, OX2 6NN, UK. +44-1865-283296, james.cummings@oucs.ox.ac.uk

Bio: Dr James Cummings helps to manage a team of developers in the Oxford University Computing Services working on digital humanities projects. He is a member of the TEI@Oxford group providing training and support for TEI projects and has served on the TEI Technical Council since 2004. He is the elected director of the Digital Medievalist project.

Sebastian Rahtz, Oxford University Computing Services, 13 Banbury Road, Oxford, OX2 6NN, UK. +44-1865-283431, sebastian.rahtz@oucs.ox.ac.uk

Bio: Sebastian Rahtz manages the Information and Support Group inside the Oxford University Computing Services. He is is a member of the TEI@Oxford group providing training and support for TEI projects, was a member of the TEI Board of Directors from 2000 to 2009, and has been a member of the TEI Technical Council since 2001. He was lead architect for the ODD system in TEI P5, and has written much of the software which underpins the TEI's work.

4. Previous venues

Although substantially modified, much of this material has been tested out for a much smaller group in an Understanding ODD pre-conference workshop for the TEI Members' Meeting and Conference 2010.

The tutors have taught TEI on summer schools at Oxford for the last 5 years to groups of c.20 delegates, and regular teach TEI concepts to small groups.

Natural Language Processing Tools for the Digital Humanities

Manning, Christopher
manning@stanford.edu
Stanford University

Large and ever-increasing amounts of text are now available digitally from many sources. Beyond raw text, there are also increasing troves of text annotated with various kinds of metadata and analysis. This data provides new opportunities in the humanities to do different kinds of analyses and at different scales, some of which blur the boundaries between the traditional analytical and critical methods of the humanities versus empirical and quantitative approaches common in the social sciences. Since texts are central to the humanities, a key opportunity is in “text mining” – making use of computers for analyzing texts, and it is here that there is much opportunity for the use of tools from Natural Language Processing. The last two decades have also seen the field of Natural Language Processing refocused on being able to process and analyze the huge amounts of available digital speech and text, partly through the use of new probabilistic and machine learning methods. This has led to the development of many robust methods and tools for text processing, many of which are within reach of the ambitious practitioner, and often are available for free as open source software.

This tutorial will survey what you can do with digital texts, starting from word counts and working up through deeper forms of analysis including collocations, named entities, parts of speech, constituency and dependency parses, detecting relations, events, and semantic roles, coreference resolution, and clustering and classification for various purposes, including theme, genre and sentiment analysis. It will provide a high-level not-too-technical presentation of what these tools do and how, and provide concrete information on what kinds of tools are available, how they are used, what options are available, examples of their use, and some idea of their reliability, limitations, and whether they can be customized. The emphasis will be at the level of what techniques exist and what you can and can’t do with them. The hope is to empower participants in envisioning how these tools might be employed in humanities research.

The rough plan of the tutorial is as follows. The plan spends a bit more time on the things that people are most likely to be able to take away and use (such as, parts of speech, NER, and parsing).

- Introduction, digital text corpora, markup, metadata, and search. Issues of spelling, tokenization and morphology (30 mins)
- Counting words, counting n-grams, collocations (20 mins)
- Part-of-speech tagging and named entity recognition (40 mins)
- Parsing: constituency and dependencies and their applications (30 mins)
- Briefer survey of methods finding more semantics: relations, events, semantic roles, and coreference resolution (20 mins)
- Clustering and classification: applications including authorship attribution, topic models, word sense disambiguation, and sentiment analysis (30 mins)
- Wrap up (10 mins)

gabmap – A Web Application for Measuring and Visualizing Distances Between Language Varieties

Nerbonne, John
j.nerbonne@rug.n
University of Groningen

Gooskens, Charlotte
c.s.gooskens@rug.nl
University of Groningen

Kleiweg, Peter
p.c.j.kleiweg@rug.nl
University of Groningen

Leinonen, Therese
t.leinonen@rug.nl
University of Groningen

Wieling, Martijn
wieling@gmail.com
University of Groningen

We frequently ask in linguistics, especially in dialectology and comparative linguistics, how similar linguistic varieties are to one another, effectively asking how similar linguistic culture is from one site to another. We operationalize the question more specifically by asking e.g. how similar the vocabulary of one variety is to another, or more interestingly how similar the pronunciations of a set of varieties are, sampled via the pronunciations of the same set of at least 30 words at a range of sites. Since there may be thousands of words and hundreds of sites, the questions must be addressed computationally. The techniques embodied in the web application have been used in dozens of scholarly papers on dialectology (see references).

At the University of Groningen the *gabmap* application has been developed that is capable of measuring differences in linguistic samples, including in particular sets of phonetic (or phonemic) transcriptions, to project present the results graphically onto maps. *Gabmap* is a graphical user interface that implements not only the comparison of vocabulary or other categorical data (essentially as percentage overlap or percentage difference) but also that of pronunciations via edit distance. Because the software is implemented as a web application users are not required to download it nor to keep it up to date by following releases. It is fairly user friendly and easily accessible and therefore enables experimentation with different techniques popular among linguists from various fields, especially dialectology and variationist linguistics.

During the workshop we will give some theoretical background about dialectometry followed by a tutorial where the theory is put into practice with exercises showing how to use the web-application. We have given similar courses in dialectology previously, for example during the Linguistic Society of America *Linguistics Institute* in 2005 at MIT and to the special meeting of the Forum *Sprachvariation* of the *Internationale Gesellschaft füt deutsche Dialektologie* in Erlangen in Oct. 2010 (www.sprachwissenschaft.uni-erlangen.de/tagung/programm.shtml). The workshop proposed here will be like the second in that it will include hands-on sessions.

The workshop will be structured as follows:

1. Introduction to dialectometry
2. Data entry: uploading dialect data, creating and uploading maps
3. Data inspection: data distribution and error detection
4. Measuring linguistic distances
5. Graphical presentations of linguistic distances: dialect maps
6. Statistical analyses: multidimensional scaling and clustering
7. Data mining, identifying influential individual variables (words, pronunciation variants)

We have named the *gabmap* collaborators as co-authors of the tutorial, but only Nerbonne and maximally one other will offer the tutorial. We can accommodate up to 20 participants.

We add a note to potential participants from non-linguistic fields. In theory one might ask the same questions of non-linguistic culture that we ask of linguistic culture, namely to what degree is e.g. the material culture of one settlement similar to that of another. We suspect that one might attack the non-linguistic question using techniques similar to the ones we will demonstrate during this tutorial, i.e. one might gather question as, but the point is purely theoretical so far, although we would welcome the chance to examine the question in a data-intensive way. If such studies are carried out, we suspect that at least the mapping facilities we demonstrate in this tutorial will be useful.

References

Alewijnse, B., Nerbonne, J., van der Veen, L. & Manni, F. (2007). 'A Computational Analysis of Gabon Varieties'.

Proceedings of the RANLP Workshop on Computational Phonology. In P. Osenova et al. (eds.). BorovetzPp. 3–12.

Gooskens, C. & Heeringa, W. (2004). 'Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data'. *Language Variation and Change.* 3: 189–207.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. thesis, University of Groningen*.

Kessler, B. (1995). 'Computational dialectology in Irish Gaelic'. *In Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics.* Dublin: EACLPp. 60–67.

Leinonen, Therese (2008). 'Factor Analysis of Vowel Pronunciation in Swedish Dialects'. *International Journal of Humanities and Arts Computing.* 2(1-2): 189-204.

Nerbonne, J. (2009). 'Data-driven dialectology'. *Language and Linguistics Compass.* 3(1): 175–198.

Nerbonne, J. & Siedle, C. (2005). 'Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede'. *Zeitschrift für Dialektologie und Linguistik.* 72(2): 129–147.

Prokic, J., Nerbonne, J., Zhobov, V., Osenova, P., Simov, K., Zastrow, T. & Hinrichs, E. (2009). 'The Computational Analysis of Bulgarian Dialect Pronunciation'. *Serdica Journal of Computing.* 3(3): 269–298.

Spruit, M. (2006). 'Measuring syntactic variation in Dutch dialects'. *Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation [Nerbonne, J., Kretzschmar, W. (eds)].* , pp. 493–506.

Yang, C. & Castro, A. (2008). 'Representing Tone in Levenshtein Distance'. *International Journal of Humanities and Arts Computing.* 2(1-2): 205–219.

Introduction to Text Analysis With Voyeur Tools

Sinclair, Stéfan
sgsinclair@gmail.com
McMaster University

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

1. Description

Are you interested in using computing methods to analyze electronic texts? Geoffrey Rockwell (University of Alberta) and Stéfan Sinclair (McMaster University) will run a hands-on workshop on using the web-based Voyeur Tools text analysis environment (voyeurtools.org). Participants can follow along with example documents or use their own. Voyeur Tools is the latest text analysis web-based system developed by TAPoR collaborators and it brings together visualization and concording tools in a fashion that allows multipanel interactive analysis or single tool analysis. Voyeur Tools runs on a high performance computing cluster and is capable of scaling to handle multiple documents and larger texts than previous web based tools.

The workshop will provide:

- An introduction to basic text analysis concepts and techniques (independent of the tool set being used)
- An introduction to different ways of using Voyeur Tools. Voyeur can be used in a multi-panel view where the different tools interact or as individual tools. Users will be shown different ways of running Voyeur Tools and how to manage panels.
- Understanding the Voyeur display. Voyeur provides a number of different panels with information from a summary of the corpus to distribution graphs. Participants will be taken through the different panels and the capabilities of each one.
- Using Voyeur Recipes for analysis. Participants will be introduced to the Voyeur Recipes, which are tutorials on how to use Voyeur for research tasks. We will start by looking at how Voyeur can be used to explore a theme through a text. We will then look at using Voyeur for diachronic study of a collection of documents over time.

- Quoting Voyeur results. Users will be introduced to Voyeur's ability to produce HTML fragments that can be used to quote results in other online documents. With Voyeur you can export your results in various ways, one of which is placing live panels into blogs or wikis.
- Integrating Voyeur into remote sites. We have developed specialized plugins that integrate with frameworks such as WordPress, Drupal and OJS. Participants will learn about these as well as how to integrate Voyeur into almost any site with a generic plugin module.

2. Instructors

Stéfan Sinclair is an Associate Professor of Multimedia. His areas of interest include computer-assisted literary text analysis, experimental visualization interfaces, and 20th Century French literature (especially Oulipo). He is the creator or co-developer of online Digital Humanities tools such as Voyeur Tools, the TAPoR Portal, the Humanities Visualization Project.

Geoffrey Rockwell is a Professor of Philosophy and Humanities Computing at the University of Alberta, Canada. He has published and presented papers in the area of philosophical dialogue, textual visualization and analysis, humanities computing, instructional technology, computer games and multimedia.

An Introduction to XForms for Digital Humanists: How XForms Can Help Your Project

Sperberg-McQueen, Michael
cmsmcq@acm.org
Black Mesa Technologies LLC, United States of America

This tutorial will introduce participants to XForms, viewed as a technology for building customized editors for XML documents.

Originally developed as a replacement for conventional HTML forms, XForms is designed to work well in the Web browser, allowing forms to be specified using the familiar facilities of XHTML and CSS. XForms is based on the model/view/controller architecture, and because the model being operated upon consists of a set of XML documents, XForms provides a convenient basis for developing custom tools for working with XML documents. With XForms, it becomes feasible for projects to develop vocabulary- and task-specific editors for use within the project. For suitably chosen tasks, specialized tools of this kind can require less training and provide better task support than full XML editors; it is thus easier to allow domain experts to examine and modify XML encoding, and routine tasks can be performed more quickly and reliably.

Any project making systematic use of XML will involve a number of tasks requiring systematic changes to XML documents, which may be more easily handled using XForms than with other techniques. For example:

- The samples in the language corpus have been divided into sentences using a probabilistic recognizer for sentence boundaries; we need to check to make sure that all the boundaries proposed by the software are in fact real sentence boundaries, and that no sentence boundaries have been missed.
- The text has been processed by a named-entity recognition engine and phrases have been marked up as personal names, clan names, place names, names of organizations, and other names, but the process by which this has happened is not fully trusted; we would like a specialist in the period and genre to check that the distinctions among place names, clan names, and personal names have been drawn correctly.

- Each of the seven hundred short documents in the project must be checked for conformance to the project's new rules for hyperlinks, and problems need to be flagged (not fixed, just flagged for later re-work).
- In a software development project, we have developed a set of XML-encoded test cases which the software should be able to handle. We have now realized that each test case needs to include metadata of a kind not originally foreseen. It needs to be added manually (or perhaps we can write a program that will get most of it right, but need a manual check to catch errors, in cases where manual fixing is cheaper than making the automatic process do the right thing).
- We are converting a few thousand MARC records into TEI headers. The conversion program tries to strip off the trailing punctuation added in the MARC record (additional full stops, semi-colons, colons, etc., depending on the internal structure of the MARC fields), but in a few cases out of every thousand fields, the algorithm is not quite right. The program writes out the TEI header with its best guess at the correct content of the element, but it also writes out the original trailing punctuation, in a specially marked processing instruction. We wish for some reasonably alert human being to go through the material and say, for each field, whether to accept the program's proposal or to restore the original punctuation. In a very small number of cases, neither the program's best guess nor the original punctuation is correct, and the reviewer must specify some third form of the data.

In each of these cases, a full-scale general-purpose XML editor can be used but has a number of drawbacks. The user performing the task must be trained in the full editor, and once the document is open in a general-purpose editor there are no limits to what changes can be made, or in cases of inattention or confusion, no limits to the damage that can be infliced on the document. It would be preferable to perform such specialized XML modification tasks in what are sometimes called "padded cell" editors, which provide simple limited interfaces for performing specific tasks. In a padded cell editor for sentence-boundary correction, for example, the user should be able to open a document, delete sentence boundaries (joining adjacent s elements), insert them (splitting s elements), save the document, and quit. An editor which provides ONLY those operations will be a lot easier to learn than a general purpose editor, and allows a careless or hapless user to do much less harm. Special-purpose editors can also incorporate knowledge of the underlying markup language and its usage in a particular project more effectively than is possible for general-purpose editors.

It has rarely been feasible, however, to use conventional programming tools to build special-purpose editors for individual XML vocabularies, let alone such specialized tasks: using standard libraries, such editors would run to thousands or tens of thousands of lines of Java or Objective C.

XForms (and some related technologies) change the equation.

XForms is built around the model / view / controller idiom, in which the model is a set of XML documents, the view is specified using XHTML and the XForms widget set, and the controller takes the form of declarative rules linking widgets to elements and attributes in the markup. That is to say: XForms can be regarded as a technology for building padded-cell editors; it has great potential for extensive application in any project making systematic use of XML.

The tutorial will comprise four blocks of material, of about 45 minutes each.

1. Introduction
   - origin and design goals of XForms
   - the XForms model/view/controller processing model
   - XForms and padded-cell editors
2. Atomic values
   - the standard XForms widgets
   - datatypes and type awareness
   - auto-calculation
   - validation in the client
   - selective display of information
   - dynamic labels
   - multi-lingual interfaces
3. More complex interfaces
   - tabbed interfaces for multi-part forms
   - repetitions
   - XForms and mixed content
4. Conclusion
   - extensions of and alternatives to XForms
   - issues in the deployment of XForms

Integrating Digital Humanities Projects into the Undergraduate Curriculum

Tomasek, Kathryn
ktomasek@wheatonma.edu
Wheaton College

Davis, Rebecca Frost
rdavis@nitle.org
National Institute for Technology in Liberal Education

Digital methods of analysis exert growing influence on the practice of many disciplines in the humanities and social sciences, yet students majoring in non-science disciplines often have little exposure to computational thinking. Although digital scholarship has become more pervasive among humanists, we have yet to recognize fully the value that collaboration with undergraduates can bring to projects in this field.

The aim of this workshop is to invite digital humanists to work together in considering how to integrate digital scholarship into undergraduate or general introductory level graduate courses. Potential motivations include:

1. advancing digital humanities within the academy, especially at the undergraduate level,
2. linking scholarship and teaching to move forward a faculty member's own project,
3. engaging students in humanities research through technology,
4. developing digital literacy to help students function well as citizens in the twenty-first century.

This workshop will present strategies for effectively integrating digital projects into undergraduate courses. By examining cases of assignments linked to digital projects, participants will consider how to make room for such assignments in a syllabus, how to tie digital projects to a course's learning outcomes, and how to scaffold both technological and content learning to allow students to make positive contributions to a project external to the course. Participants will leave with a set of pedagogical strategies for thinking about digital projects, preliminary plans for assignments for their own courses, and suggestions for how to find collaborative partners in library and technology services for such projects on their home campuses.

Part One: Rebecca Frost Davis will present an overview of ways that digital humanities projects have been integrated into the curriculum, contextualized through the pedagogical approach of problem-based learning and the principles of liberal education. Participants will look at a variety of examples, including the Homer Multitext Project (`www.homermultitext.org`) , SmartChoices (`smartchoices.trincoll.edu`) , and the NINES Collex (`www.nines.org`) .

Part Two: Kathryn Tomasek will present a case study from the Wheaton College Digital History Project. A transcription and encoding assignment, this module can be dropped into multiple courses. It includes scaffolded assignments; a teaching collaboration that involves a faculty member, a technologist, and an archivist; and multiple opportunities for students to create and use new historical data whilst contributing to larger digital history projects.

Part Three: Participants will brainstorm and workshop assignments for their own courses in breakout groups. Discussion will center on three areas, practical questions about how to integrate assignments into a course, how to pace and scaffold work on digital projects, and questions about collaborative pedagogy. The workshop will conclude with discussion of individual assignment ideas from the small groups.

2. Presenters

The organizers bring to the workshop fifteen years' combined experience in integrating technology projects into the undergraduate curriculum. Rebecca Frost Davis, Program Officer for the Humanities at the National Institute for Technology in Liberal Education (NITLE), has been teaching faculty development workshops on the effective pedagogical application of technology since 2002. Currently, she heads NITLE's initiative in digital humanities and researches the growth of the field at small liberal arts colleges. Kathryn Tomasek, Associate Professor of History at Wheaton College in Massachusetts, has been using the Text Encoding Initiative in her courses since fall 2004. She is Co-Director of the Wheaton College Digital History Project, a long-term digitization project that has employed students as summer research assistants and now includes a transcription and markup module that Tomasek uses in multiple courses for advanced-level History majors. Davis and Tomasek will offer the workshop described here as a Bootcamp at THATCamp LAC, June 4-5, 2011, at St. Norbert College in De Pere, Wisconsin.

References

Blackwell, C., Martin, T. R. (2009). 'Technology, collaboration, and undergraduate research'. *Digital Humanities Quarterly.* 3(1). http://www.digitalhumanities.org/dhq/vol/3/1/000024/000024.html.

Cavanagh, S. (2010). 'Bringing our brains to the humanities: increasing the value of our classes while supporting our futures'. *Pedagogy.* 10(1): 131-142. http://muse.jhu.edu/journals/pedagogy/v010/10.1.cavanagh.html.

Network and Topical Analysis for the Humanities using NWB and Sci2

Weingart, Scott
scbweing@indiana.edu
Indiana University

Börner, Katy
katy@indiana.edu
Indiana University

Duhon, Russell
rduhon@indiana.edu
Indiana University

Linnemeier, Micah
mwlinnem@indiana.edu
Indiana University

Phillips, Patrick
pataphil@gmail.com
Indiana University

Biberstine, Joseph
jrbibers@indiana.edu
Indiana University

Tank, Chintan
tankchintan@gmail.com
Indiana University

Kong, Chin Hua
kongch@indiana.edu
Indiana University

1. Abstract

More and more, research in the humanities requires making use and sense of datasets that represent the structure and dynamics of complex natural and man-made systems. Recent trends in the digital humanities have resulted in the wide-scale availability of this data. The analysis, navigation, and management of these large-scale, dynamically changing datasets requires a new kind of tool, a macroscope (from macro, great, and skopein, to observe).

Microscopes empowered our naked eyes to see cells, microbes, and viruses, thereby advancing the progress of biology and medicine. Telescopes opened our minds to the immensity of the cosmos and prepared mankind for the conquest of space. Macroscopes

promise to help us cope with another infinite: the infinitely complex. They allow us to detect patterns, trends, and outliers, give access to details, present a 'vision of the whole,' and assist our 'synthesis' of what we observe. While most microscopes and telescopes are static physical instruments, macroscopes are continuously changing bundles of software deployed as cyberinfrastructures, Web services, or standalone tools.

This tutorial presents and demonstrates CIShell powered tools such as the Science of Science (Sci2) Tool (http://sci.slis.indiana.edu/sci2) and the Network Workbench (NWB) Tool (http://nwb.slis.indiana.edu). The NWB Tools is a network analysis, modeling, and visualization toolkit for physics, biomedical, social science, and other multidisciplinary research. The Sci2 Tool was specifically designed for researchers and science policy makers interested to study and understand the structure and dynamics of science. These versatile tools can be utilized for humanities data, allowing humanists to explore topical or social networks within their areas of study. They will be utilized for the analysis of temporal, geospatial, topic, and network datasets, and the professional visualization of analysis results by means of large-format charts and maps. Both tools are standalone desktop applications that install and run on Windows, Linux x86 and Mac OSX.

2. Outline

15 Min.Marcoscope Design and Usage

45 Min. Sci2 Tool Basics

- Download and run the tool.
- Load and clean a dataset using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

15 Min. Sci2 Workflow Design. Padgett's Florentine Families - Prepare, load, analyze, and visualize family and business networks from 15th century Florence.

15 Min. Break

30 Min. Sci2 Research Demonstration I. Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

30 Min. Sci2 Research Demonstration II. The Republic of Letters - Find central correspondents in Early-Modern Europe.

30 Min. Q&A and Technical Assistance

3. Software Needed

Custom software that uses Java 1.5 or higher and the OSGI/CIShell (http://cishell.org) core together with algorithm plugins and sample datasets from the Network Workbench Tool (http://nwb.slis.indiana.edu) and Science of Science Tool (http://sci.slis.indiana.edu).

4. Suggested Reading

- Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), ARIST, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, Volume 37, Chapter 5, pp. 179-255. http://ivl.slis.indiana.edu/km/pub/2003-borner-arist.pdf
- Börner, Katy, Sanyal, Soma and Vespignani, Alessandro (2007). Network Science. In Blaise Cronin (Ed.), ARIST, Information Today, Inc./ American Society for Information Science and Technology, Medford, NJ, Volume 41, Chapter 12, pp. 537-607.http://ivl.slis.indiana.edu/km/pub/2007-borner-arist.pdf
- Scott Weingart, Hanning Guo, Katy Borner, Kevin W. Boyack, Micah W. Linnemeier, Russell J. Duhon, Patrick A. Phillips, Chintan Tank, and Joseph Biberstine (2010) Science of Science (Sci2) Tool User Manual. Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington. http://sci.slis.indiana.edu/registration/docs/Sci2_Tutorial.pdf

5. Instructors

Scott Weingart is a doctoral student at Indiana University studying History of Science and Information Science. His primary research is on the Republic of Letters in Early Modern Europe, and has worked with the CKCC project in the Netherlands and Dr. Robert A. Hatch at the University of Florida on digitizing, visualizing, and analyzing early modern correspondence networks. Scott focuses on the intersection of computational analysis and the humanities, and how each can shape the other.

6. Previous Venues

NWB and Sci2 have been presented at dozens of previous venues with audience sizes ranging from under 10 to over 100. A full list of our previous presentations can be found at http://cns.slis.indiana.edu/presentations/. The most recent workshop presented specifically for humanists was at the NEH-funded Networks and Network Analysis for the Humanities Summer Institute at UCLA in August 2010. There were approximately 40 audience members in attendance, and the workshop resulted in over half of the attendees using the Sci2 tool in their final presentations for the institute.

Panels

Virtual Cities/Digital Histories

Allen, Robert C.
rallen@email.unc.edu
University of North Carolina at Chapel Hill

Smith, Natasha
nsmith@email.unc.edu
University of North Carolina at Chapel Hill

Lach, Pamella
plach@email.unc.edu
University of North Carolina at Chapel Hill

Marciano, Richard
richard_marciano@unc.edu
University of North Carolina at Chapel Hill

Speed, Chris
c.speed@eca.ac.uk
Edinburgh College of Art

Presner, Todd
presner@humnet.ucla.edu
UCLA

Ethington, Philip
philipje@usc.edu
Univ. of Southern California

Shepard, David
shepard.david@gmail.com
UCLA

Hou, Chien-Yi
chienyi@unc.edu
University of North Carolina at Chapel Hill, United States of America

Johanson, Christopher
cjohanson@gmail.com
UCLA

---

1. Going to the Show and Main Street, Carolina

Robert C. Allen, Natasha Smith, Pamella Lach; University of North Carolina at Chapel Hill

`http://www.docsouth.unc.egu/gtts`

`http://www.http://mainstreet.lib.unc.edu/`

Going to the Show documents and illuminates the experience of movies and moviegoing in North Carolina from the introduction of projected motion pictures (1896) to the end of the silent film era (circa 1930). Through its innovative use of more than 1000 digitally stitched and georeferenced Sanborn® Fire Insurance maps of forty-five towns and cities between 1896 and 1922, the project situates early moviegoing within the experience of urban life in the state's big cities and small towns. Supporting its documentation of more than 1300 movie venues across 200 communities is a searchable archive of thousands of contemporaneous artifacts: newspaper ads and articles, photographs, postcards, city directories, and 150 original architectural drawings.

Main Street, Carolina (in development) is a digital history toolkit designed to allow cultural heritage organizations in North Carolina to preserve, document, and share the history of their downtowns by creating and managing digital content and displaying it on interactive historic maps.

2. Hypercities

Philip J. Ethington, Univ. of Southern California; Todd Presner, Christopher Johanson, David Shepard, UCLA

`http://hypercities.com/`

Built on the idea that every past is a place, *HyperCities* is a digital research and educational platform for exploring, learning about, and interacting with the layered histories of city and global spaces. Developed though collaboration between UCLA and USC, the fundamental idea behind *HyperCities* is that all stories take place somewhere and sometime; they become meaningful when they interact and intersect with other stories. Using Google Maps and Google Earth, *HyperCities* essentially allows users to go back in time to create and explore the historical layers of city spaces in an interactive, hypermedia environment. A HyperCity is a real city overlaid with a rich array of geo-temporal information, ranging from urban cartographies and media representations to family genealogies and the stories of the people and diverse communities who live and lived there.

3. T-RACES

Richard Marciano, Chien-Yi Hou; University of North Carolina at Chapel Hill

`http://salt.unc.edu/T-RACES/`

T-RACES (Testbed for the Redlining Archives of California's Exclusionary Spaces) presents Residential Security Maps created by the Home Owners Loan Corp in the 1930s for eight cities in California along with supporting documentation. These maps categorized

specific areas in cities according to four color-coded categories based on racial, ethnic and economic characteristics of residents and potential home buyers. These so-called “redlining” maps were used by local financial institutions to make home mortgage decisions and had a significant impact on the fate of urban neighborhoods for decades. The site allows users to view the maps, query a wide range of supporting data, and download KML files for use with Google Earth.

4. Walking Through Time and Tales of Things

Chris Speed, Edinburgh College of Art

`http://walkingthroughtime.eca.ac.uk/`

`http://www.talesofthings.com`

Walking Through Time is a smart phone web application that allows architectural historians, conservationists and tourists to download historical maps of Edinburgh when standing in a specific location and to annotate them. They can walk through real space whilst following a map from 200 years ago (for example) and tag and attach links to the map that offer historical and contextual information.

Tales of Things is part of a research project called TOTeM that will explore social memory in the emerging culture of the Internet of Things. Researchers from across the UK have provided this site as a platform for users to add stories to their own treasured objects and to connect to other people who share similar experiences. This will enable future generations to have a greater understanding of the object’s past and offers a new way of preserving social history. Content will depend on real people’s stories, which can be geo-located through an on-line map of the world where participants can track their object even if they have passed it on. The object will also be able to update previous owners on its progress through a live Twitter feed which will be unique to each object entered into the system.

Integrating Digital Papyrology

Baumann, Ryan
rfbaumann@gmail.com
University of Kentucky, Lexington

Bodard, Gabriel
gabriel.bodard@KCL.AC.UK
King's College, London

Cayless, Hugh
hugh.cayless@nyu.edu
New York University

Sosin, Joshua
joshua.sosin@DUKE.EDU
Duke University

Viglianti, Raffaele
raffaele.viglianti@KCL.AC.UK
King's College, London

"Integrating Digital Papyrology" (IDP) is a series of projects funded by the Andrew W. Mellon Foundation. It represents the integration of three longstanding digital papyrology efforts: the Duke Databank of Documentary Papyri (DDbDP), the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV), and the Advanced Papyrological Information System (APIS), and the migration to a single format, the EpiDoc recommendations for the application of TEI XML to ancient documentary texts. [1]

This panel aims first to present an historical overview of the transformation of the DDbDP from a digital index of print scholarship to a community-managed resource for peer-reviewed scholarly control of core disciplinary assets, and then to suggest some ways in which we hope that the new suite of tools that we have been building around the data may help to transform scholarship in this important domain.

The DDbDP began in 1983 as a collaboration between Duke University and David R. Packard. Greek and Latin were encoded in Beta Code and searchable on CD using custom software. In 1996/7 the DDbDP migrated its authoritative version to the web-based Perseus Project, encoded in Beta Code and TEI SGML. In 2004/5 the DDbDP and HGV began mapping their theretofore discrete but complementary sets of text and descriptive metadata, while Duke obtained a planning grant from the Mellon Foundation.

Papyrologists, IT specialists, librarians and university administrators came together to map out a sustainable future for the DDbDP. The way forward was clear: open source, standards-based development, greater collaboration, increased vesting of data-control in the user community, and greater interoperability with other projects.

With clear objectives and generous support from the Mellon Foundation, Duke launched IDP1 (2007/8). Its goals were to migrate the DDbDP from SGML to EpiDoc XML, and from Betacode Greek to Unicode Greek; to merge DDbDP texts and HGV metadata and translations into a single stream; to map these texts to corresponding APIS records, including images; to enhance the Papyrological Navigator (PN—see http://papyri.info) to enable search of the newly merged data. The fruits of these efforts were released under open access provisions in October 2008 (all content under CC BY and software under GNU GPL). In October 2008 the team began work on IDP2, again with support from Mellon. At the same time, APIS, under a grant from the National Endowment for the Humanities, began work on enhancing the PN, and the two projects joined forces. The results of IDP2 are (1) improved usability of the PN search interface on the merged and mapped data from the DDbDP, HGV and APIS, (2) facilitated third-party use of the data and tools, and (3) a version controlled, transparent and fully audited, multi-author, web-based, real-time, tagless editing environment (SoSOL), which—in tandem with a new editorial infrastructure—will allow the entire community of papyrologists to take control of the process of populating these communal assets. The discipline is now on the cusp of having the the entire life-cycle of the papyrological discipline represented online in a transparent, open, peer-reviewed and community-driven environment.

References

Cayless, Hugh (2009). 'Epigraphy in 2017'. *Digital Humanities Quarterly.* 3.1. `http://www.digitalhumanities.org/dhq/vol/3/1/000030/000030.html`.

Notes

1. `http://epidoc.sourceforge.net/`; see also Cayless (2009) (`#cayl2009`)

PAPER 1

Lessons from the conversion of the Duke Databank of Documentary Papyri from legacy formats into EpiDoc TEI XML

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

Sosin, Joshua
joshua.sosin@duke.edu
Duke University

Viglianti, Raffaele
raffaele.viglianti@kcl.ac.uk
King's College London

When the DDbDP began in 1983 at Duke University, [1]Greek and Latin were encoded in Beta Code (an ASCII representation of text in different scripts and of various sigla and structural features, also used by the TLG,[2]that combines language encoding and some semantic markup features) and searchable on CD-ROM via a dedicated platform provided by the Packard Humanities Institute. These texts were entered manually by students at Duke, following a data entry manual, and using published editions of papyrological texts as the basis. When the DDbDP migrated from the CD-ROMs to the Web-based Perseus Project, the texts were machine-translated to a format mixing Beta Code and TEI P3 SGML (and the resultant hybrid encoding formed the basis of an updated data entry manual). [3]

The first stage of the Mellon-funded Integrating Digital Papyrology project (2007-2008) involved the conversion of the DDbDP from largely regular and sometimes consistent format, but highly varied contents, into Unicode and EpiDoc XML. At this point the hybrid encoding scheme was a mixture of Beta Code (varying slightly over thirteen years of evolution and multiple generations of graduate students), machine-tagged TEI SGML (and hand-entered since the Perseus conversion). This basic consistency, compromised but not destroyed by this technical history of the project, is further complicated by the wide variety of papyrological material, including editions published over a century by editors with differing standards of editorial detail, and even

occasionally inconsistent uses of conventions. These variations were somewhat, but not entirely, flattened by data entry practice.

Due to the size of the dataset, some 55,000 records, the conversion of these texts into both Unicode character encoding and EpiDoc-conformant TEI XML structural and semantic features had to be almost entirely automated, with only a small amount of human intervention for difficult or unique cases. This first round of conversion work was carried out by a team at the Centre for Computing in the Humanities at King's College London. The tools to convert from the complex, legacy formats to the more sustainable, open standard TEI XML were open source and based on ongoing work in the EpiDoc community.

The CCH team built the data conversion as a four-step pipeline: (1) the SGML was turned into validating XML, and entities resolved, using a tool based on OSX; [4](2) the Beta Code representation of Greek and Coptic characters and certain symbols was converted to Unicode using Transcoder; [5](3) a new set of regular expressions was added to CHETC, [6]a regular expression-based conversion tool to convert much of the structural Beta Code to XML; (4) XSLT was applied to transform the XML produced by steps 1 and 3 into validating EpiDoc. In order to write these steps simultaneously, and iteratively to improve the tools, the process was pipelined so that the master copy of the texts remained the Beta Code/SGML hybrid, which continued to be manually improved. Only at the end of the process, when the output validated to the EpiDoc DTD, would the transformation be complete, and hand-fixes begin to be performed on this new version of the DDbDP.

In the second phase of the project, work on the DDbDP focussed on ongoing hand-fixes to the now canonical XML to improve the consistency of the text. This became ever more essential as the SoSOL tool (see Baumann, below) requires very consistent, valid EpiDoc. The EpiDoc recommendations were also updated to TEI P5 in this phase, and both Duke texts and HGV metadata encoded in this new format. The third phase (awaiting funding decision) will work to make the metadata translation into EpiDoc stable, rather than a pipelined crosswalk as it is currently.

Among the important lessons we learned from the DDbDP translation from the hybrid format to standarized TEI XML are the following technical points (all of value not only to our ongoing work with HGV metadata, but potentially to others contemplating large-scale up-coding from legacy formats):

1. Beta Code was a well-designed encoding scheme for Ancient Greek script, but even so there were issues that needed to be resolved. Where Unicode has two characters for medial and terminal Greek sigma (σ and ς), Beta considers these to be unambiguous by position and so uses 'S' to encode both. For a converter that had to parse mixed-content XML, relying on position proved unreliable. Similarly, Unicode now has separate ranges for Greek and Coptic alphabets, but Beta only distinguishes between the scripts in the case of the six characters that are unique to Coptic. In the Classics world there are also tools in wide use that expect Beta Code input (for example the Morpheus morphological parser), which is no doubt part of the reason why this legacy format is in use so long after the general adoption of Unicode.
2. Because of the complexity of even what seemed to be the simplest task, above, the integrity of the processing pipeline was especially important, since changes at every stage could otherwise turn out to be irreversible. At one stage, before the end of the project, we decided to abandon the first step of the conversion process and consider the validating XML version to be the master copy. It was only after several rounds of hand-fixes and global corrections had been made to this new master copy, that we noticed an error in the first step of the conversion process, which lost an important distinction in some thousands of cases. It was possible, but difficult, to go back and diff the new version against the SGML in this case, but the issue remains that moving from a legacy master text to an automated improvement of same requires careful checking.
3. Possibly the most interesting challenge in this conversion process was the hybrid data format that combined Beta Code sigla, SGML tags and entities to encode structural and semantic information about the papyrological editions. Translating these multiple encoding schemes to valid EpiDoc XML was not as simple as writing transformations for each of the levels, since a span of lost or restored text (represented by Beta Code brackets) and an abbreviation (represented by SGML tags) might overlap, leading to ill-formed XML. (In fact, it is probably because of this thorny problem of overlapping spans that the Beta Code was never fully translated to SGML in the initial conversion). The problem was ultimately solved by a combination of fine-tuned Regex, XSLT and some hand-correction.
4. The interaction of regular expressions and XSLT in the translation of the markup was also an important

lesson. It was obvious from the start that Regexes were suited to replacing textual sigla with XML, and XSLT to transforming SGML/XML tags into other XML, but the relationship between these two processes was both complex and interesting. In some cases it was simpler to replace sigla with well-formed but invalid XML tags, and then fix those to good EpiDoc with XSLT; in other cases a combination of hand-fixing and one-off regular expressions were used to regularize inconsistencies to a single, consistent error, which the two stages of the markup translation process could then fix in one pass. The choice of which stage of the process to fix any given issue proved to be worth careful planning. (By the same token, it was important to decide when the scale of a problem meant hand-fixing was more efficient than a dynamic solution in the pipeline.)

5. Another important lesson was that sometimes the markup in the papyrological editions was ambiguous in a way that spanned two different semantic distinctions in the EpiDoc guidelines (for example, corrections and regularizations used the same encoding in the DDbDP, but are represented by choice/sic/corr and choice/orig/reg respectively in TEI). Given the scale of the project it was impossible to disambiguate the many thousands of instances of these, and markup was devised to preserve this lack of granularity.
6. Although 55,000 texts is not huge by database standards, for a corpus of ancient texts in XML, processed using XSLT it was larger than we had dealt with before. The tools that had been adequate for dealing with smaller numbers of text (individual Java calls to the Saxon processor; Subversion for version control) led to processes and commits that took up to 24 hours at a time on the scale of this project. The solutions were relatively straightforward: pipelining multiple texts to Saxon through one Java call saved 95% of the processing time; migrating to Git from Subversion improved commit speed manyfold. Such optimizations are clearly essential at large scales.
7. The management of a distributed project (including members from five institutions based in Alabama, Heidelberg, Kentucky, London, New York, and North Carolina) offered additional lessons. Project communication is handled via several channels: there is a project mailing list for asynchronous conversations, we hold weekly conversations in Skype between representatives from each location, there is an always-on chatroom on IRC that project members use for daily conversations, and we have impromptu Skype calls when voice communication is needed. In addition, we have held several week-long, all-hands meetings during which planning and intensive collaborative work is done. Project communication and coordination does not follow the management hierarchies of the member institutions, and this flat organization greatly improves efficiency and implementation speed.

References

Nicholasm, Nick et al. (1999-2000). *The TLG Beta Code Manual,.* `http://www.tlg.uci.edu/encoding/BCM2010.pdf`.

Oates, John (1993). 'The Duke Databank of Documentary Papyri'. *Accessing Antiquity: The Computerization of Classical Studies.* Jon Solomon (ed.). Arizona, pp. 62-72.

Sosin, Joshua (2010). 'Digital Papyrology'. *Congress of the International Association of Papyrologists.* Geneva, 19 August 2010. `http://www.stoa.org/archives/1263`.

Notes

1. Oates (1993)
2. Nicholas (1999)
3. Sosin (2010).
4. OSX, OpenJade Distribution, available: `http://openjade.sourceforge.net/doc/sx.htm`.
5. Hugh Cayless, Transcoder, available: `http://epidoc.sourceforge.net/resources.shtml#transcoder`.
6. Hugh Cayless, Chapel Hill Electronic Text Converter, available: `http://epidoc.sourceforge.net/resources.shtml`.

PAPER 2

The Papyrological Navigator: Project Integration with RDF

Cayless, Hugh
hugh.cayless@nyu.edu
New York University

One of the major difficulties encountered during IDP1 was the problem of aggregating related content. The main purpose of the Papyrological Navigator (PN) is to present a merged, searchable view of the DDbDP, HGV, and APIS datasets. IDP1 attempted to produce an aggregated dataset, where related DDbDP and

HGV records were merged into a single EpiDoc XML document. The mapping processes whereby this was accomplished were flawed, however, and resulted in both spurious and missing relationships. In addition, combining documents from projects with different update schedules and sometimes different interpretations of what constitutes a document was problematic. For IDP2, it was clear that the mapping process would have to be re-imagined. The new process, described in this paper, uses a combination of semantic web technology and standard tools to produce a merged collection.

IDP draws on datasets from the DDbDP, HGV, APIS, and Trismegistos [1]. The varied origins of these projects meant that different decisions were made in each about how to record information about source documents. For example, if several pieces of papyrus contain text that are part of the same "document," then DDbDP will tend to transcribe the whole, marking the separate sheets of papyrus with elements. HGV on the other hand, tends to treat each piece as an entity, each with its own record. APIS, since it is essentially a unified export of the catalogs of museums and libraries with papyrus holdings, treats the document the way the host collection does. While APIS is organized according to the source collection, DDbDP and HGV use the publication history of the papyrus as an organizing principle. HGV has the idea of a "principal edition," i.e. the best published version of the papyrus. DDbDP is similarly based on a single published edition, though HGV and DDbDP may not agree on what is the principal edition. When a better publication comes out for an existing document, DDbDP replaces the existing document with the new version, leaves a "stub" record with a pointer to the new version in place of the old, and the new edition points back at the old one. HGV, similarly, sometimes retires identifiers. But these updates are not always efficiently propagated across projects, so references may become stale over time.

DDbDP identifiers are constructed from short-form citations of the document edition on which the transcription was based, so, for example bgu;1;2 is the identifier for the 2nd item in the first volume in the BGU (*Aegyptische Urkunden aus den Königlichen (later Staatlichen) Museen zu Berlin, Griechische Urkunden*) series. Trismegistos assigns numeric identifiers to individual documents (e.g. 1234), and HGV bases its identifiers on these. When HGV data is finer-grained than TM, it will append alphabetic characters to the TM id (e.g. 1234a, 1234b, etc.). APIS bases its identifiers on the collection name plus a number, e.g. berkeley.apis.15.

DDbDP and HGV began the process of reconciling their collections in 2004, so where there are correspondences, the documents include the id numbers of related records in their document headers. For example, http://papyri.info/ddbdp/p.oxy;4;744/ source encodes identifiers both for itself and for the related HGV record in its publicationStmt:

```
<publicationStmt>
  <authority>NYU Digital Library Technology Services</authority>
  <idno type="filename">p.oxy.4.744</idno>
  <idno type="ddb-perseus-style">0181;4;744</idno>
  <idno type="ddb-hybrid">p.oxy;4;744</idno>
  <idno type="HGV">20442</idno>
  <idno type="TM">20442</idno>
  <availability>
    <p>© Duke Databank of Documentary Papyri. This work is licensed under a <ref
      type="license" target="http://creativecommons.org/licenses/by/3.0/">Creative Commons
      Attribution 3.0 License</ref>.</p>
  </availability>
</publicationStmt>
```

Likewise, `http://papyri.info/hgv/20442/source`, the related HGV record, contains the DDbDP identifier in its header:

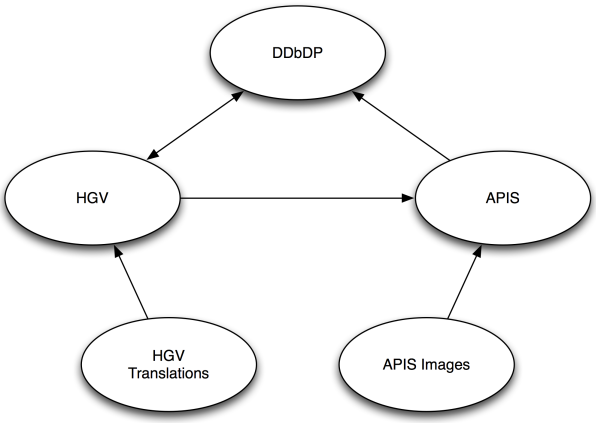
```
<publicationStmt>
  <idno type="filename">20442</idno>
  <idno type="TM">20442</idno>
  <idno type="ddb-perseus-style">0181;4;744</idno>
  <idno type="ddb-filename">p.oxy.4.744</idno>
  <idno type="ddb-hybrid">p.oxy;4;744</idno>
</publicationStmt>
```

In this (simple) case, the relationships are easy to establish, and can be represented in RDF (using N3 syntax) thus: `<http://papyri.info/ddbdp/p.oxy;4;744/source> <http://purl.org/dc/terms/relation> <http://papyri.info/hgv/20442/source>` `<http://papyri.info/hgv/20442/source> <http://purl.org/dc/terms/relation> <http://papyri.info/ddbdp/p.oxy;4;744/source>` That is, the resource http://papyri.info/ddbdp/p.oxy;4;744/source has a "relation" (which is a term drawn from the Dublin Core Metadata Initiative Terms [2]list) http://papyri.info/hgv/20442/source, and the latter has a reciprocal relationship to the former.

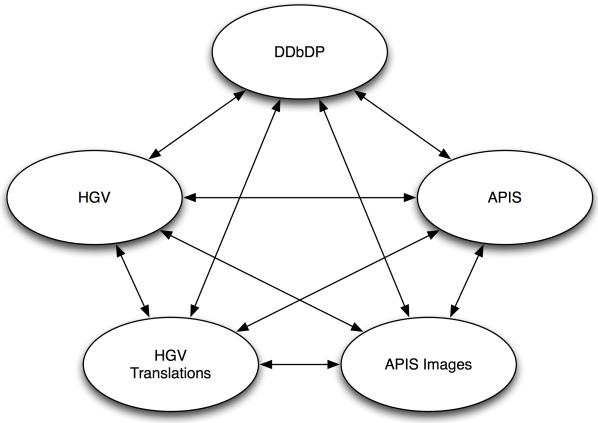
The question of aggregation becomes much more complicated when APIS enters the equation. Because of its origin as a union catalog (the data for which was contributed by a variety of insitutions), the quality of APIS's information is more uneven. The method used by the first version of the PN to establish relationships with DDbDP, namely looking for matching citations, produced results that were sometimes inaccurate. For IDP2, we looked for pieces of information that could be extracted (like the <idno>s in DDbDP and HGV) and represented as RDF triples. HGV contains URLs for images of papyri and ostraka when they are available, and when these are hosted at Columbia (the original home of APIS), they contain the same APIS id as the APIS documents do. In addition,

APIS records sometimes include references to the relevant DDbDP document in their bibliography: `<bibl type="ddbdp">P.Oxy.:4:744</bibl>` This does not use the DDbDP identifier scheme, but it can be converted to the DDbDP identifier, p.oxy;4;744 easily enough. APIS contains some spurious references of this type (to collections that don't exist, for example), but all that is needed to check their validity is to check for the existence of a corresponding DDbDP file.

In addition to the collections mentioned above, there are also HGV translations, which use the same numbering scheme as the HGV metadata, and APIS images, which are represented by a static RDF file mapping image names to APIS ids. A representation of the incomplete graph that can be constructed by extracting these relations from the source documents looks something like:



So there are locations in an IDP EpiDoc document (and in associated artifacts) where we can check for reference information, and from these we can build a (partial) graph of relationships between documents in the three collections. The database import method is simple. A program crawls each directory in our repository containing EpiDoc documents, runs each through an XSLT transformation which converts it to RDF XML, and inserts that RDF into a Mulgara# triple store. Once the relations have been converted to RDF, we can do some simple inferencing to "fill out" the rest of the incomplete relation graph. In practice, this means running SPARQL queries that produce results like "A is related to B, because B is related to A" and "A is related to C, because A is related to B and B is related to C," and then inserting the results of those queries back into the triple store. The resulting graph looks more like:



Besides relation data, we can also extract hierarchical and bibliographical information from the collection. Because DDbDP's identifiers are hierarchical and carry meaning, they can be decomposed and used to identify not just the items, but their containers, the volumes and series in which they were published. APIS ids can be decomposed to collection and item. HGV and TM ids are opaque, but HGV's principal edition bibliographic record can be used to construct a hierarchy for that collection. HGV's bibliography also points to the primary source for its record.

The APIS, DDbDP, and HGV collections are hosted in the Papyrological Navigator, while TM has its own website. This means that we are able to mint URLs using the APIS, DDbDB, and HGV identifiers, so http://papyri.info/ddbdp/bgu;1;2/source represents the EpiDoc XML document that transcribes BGU 1 2, http://papyri.info/ddbdp/bgu;1;2 retrieves an HTML document that aggregates the DDbDP transcription with metadata from HGV #8961 and links out to TM #8961, http://papyri.info/ddbdp/bgu;1;2/rdf is shorthand for an RDF query that pulls together a subgraph of triples with http://papyri.info/ddbdp/bgu;1;2/source as the subject, and so on.

The primary internal uses of the RDF triple store in papyri.info are as the driver for generating and indexing the site, and as a means for SoSOL (see Ryan Baumann's paper) to reference valid identifiers. The method for generating the site uses the triple store to load, first, all of the identifiers associated with DDbDP using the hierarchical relations, then all of the HGV ids not linked to a DDbDP identifier, then all of the unrelated APIS identifiers. The collections are then transformed to HTML, and the files corresponding to file identifiers (along with any related EpiDoc documents) are transformed both to HTML and to Solr add documents, which are then ingested into the search engine. In this fashion, we create the entire

papyri.info site, with correctly aggregated data, using a few basic tools.

Developments slated for the near future include using the triple store as a container for relations to external projects. An effort was made during the summer of 2010 to link HGV place names to Pleiades and Trismegistos Places URLs, and once the data has been processed, it will be loaded into our triple store and used as the basis for linking to data at these sites. We hope, for example, to be able to display maps drawn from Pleiades, as well as linking to the site. Another planned use for the triple store is to have it handle the normalization and display of bibliography for papyri.info.

The method outlined above for extracting data from source documents, storing it as RDF triples, and then using inferencing to fill out gaps in the graph has potential applications well beyond papyri.info. For example, it could be used to manage relationships between EAD finding aids and digitized archives, or to merge data from other overlapping collections. Any data management situation where joins between discrete but related objects as desired, and where it is possible to extract a partial relationship graph from the sources could profit from using semantic web tools in this fashion.

Notes

1. See http://www.trismegistos.org/
2. DCMI Metadata Terms, available: http://dublincore.org/documents/dcmi-terms/.

PAPER 3

The Son of Suda Online: Collaborative editing and workflow tool

Baumann, Ryan
rfbaumann@gmail.com
University of Kentucky

The Son of Suda On Line [1](SoSOL) is one of the main components of the Integrating Digital Papyrology project (IDP), aiming to provide a repurposable web-based editor for the digital resources in the DDbDP and HGV. SoSOL integrates a number of technologies to provide a truly next-generation online editing environment. Using JRuby [2]with the Rails [3]web framework, it is able to take advantage of Rails's wide support in the web development community, as well as Java's excellent XML libraries and support. This includes the use of XSugar [4]to define an alternate, lightweight syntax for EpiDoc XML markup, called Leiden+. Because XSugar uses a single grammar to define both syntaxes in a reversible and bidirectional manner, this is ideal for reducing side effects of transforming text in our version-controlled system. SoSOL uses the Git [5]distributed version control system as its versioning backend, allowing it to use the powerful branching and merging strategies it provides, and enabling fully-auditable version control. SoSOL also provides for editorial control of changes to the main data repository, enabling the democracy of allowing anyone to change anything they choose while preserving the academic integrity of canonical published data. This talk will provide a demonstration of these features of SoSOL as implemented for IDP2, as well as a discussion of its repurposable design for applicability to other projects and the ongoing documentation work being done to increase usability and adoption in the wider community.

1.1. Next-Generation Version Control

Many online editing environments, such as MediaWiki, use an SQL database as the sole mechanism for storing revisions. This can lead to a number of problems, such as scaling (most SQL servers are not performance optimized for large text fields) and distribution of data (see for example the database downloads of the English Wikipedia, which have been notoriously problematic for obtaining the full revision history). Most importantly, they typically impose a centralized, linear, single-branch version history. Because Git is a distributed version control system, it does not impose any centralized workflow. As a result, branching and merging have been given high priority in its development, allowing for much more concurrent editing activity while minimizing the difficulty of merging changes. SoSOL's use of Git is to have one "canonical" Git repository for public, approved data and to which commits are restricted. Users and boards each get their own Git repositories which act as forks of the canonical repository. This allows them to freely make changes to their repository while preserving the version history as needed when these changes are merged back into the canonical repository. These repositories can also be easily mirrored, downloaded, and worked with offline and outside of SoSOL due to the distributed nature of Git. [6]This enables a true democracy of data, wherein

institutions still retain control and approval of the data which they put their names on, but any individual may easily obtain the full dataset and revision history to edit, contribute to, and republish under the terms of license.

1.2. Alternative Syntax for XML Editing

While XML encoding has many advantages, users inexperienced with its use may find its syntax difficult or verbose. It is still desirable to harness the expertise of these users in other areas and ease their ability to add content to the system, while retaining the semantically explicit nature of XML markup. To do this, we have used XSugar to allow the definition of a "tagless" syntax for EpiDoc XML, which resembles that of the traditional printed Leiden conventions for epigraphic and papyrological texts where possible. Structures which are semantically ambiguous or undefined in Leiden but available in EpiDoc (e.g. markup of numbers and their corresponding value) have been given additional text markup, referred to comprehensively as Leiden+. XSugar enables the definition of this syntax in a single, bidirectional grammar file which defines all components of both Leiden+ and EpiDoc XML as correspondences, which can be statically checked for reversibility and validity. This provides much more rigorous guarantees of these properties than alternatives such as using separate XSLT stylesheets for each direction of the transform, as well as encoding the relation between the components of each syntax in a single location.

As an example, we might have the following XML fragment:

```
<div xml:lang="grc" type="edition" xml:space="preserve">
  <ab>
    <lb n="1"/><supplied reason="lost">ἔτους</supplied> <supplied
      reason="lost" cert="low"><num value="1">α</num> </supplied> <supplied
      reason="lost">Αὐτοκράτορος</supplied> <gap reason="illegible" quantity="2"
      unit="character"/><gap reason="lost" quantity="1" unit="character"/><gap
      reason="illegible" quantity="2" unit="character"/>του
    <lb n="2"/><gap reason="lost" quantity="12" unit="character"
      precision="low"/> Σεβαστοῦ
    <lb n="3"/><supplied reason="lost"><expan>εἴργ<ex>ασται</ex></expan>
      <expan>ὑ<ex>πὲρ</ex> χω<ex>ματικῶν</ex></expan></supplied>
      <expan><supplied reason="lost">ἔ</supplied>ργ<ex>ων</ex></expan> τοῦ
      αὐτο<unclear>ῦ</unclear> πρώτου <expan><ex>ἔτους</ex></expan>
    <lb n="4"/><gap reason="lost" extent="unknown" unit="character"/> <num
      value="20">κ</num> <num value="26">κς</num> ἐ<supplied
      reason="lost">ν</supplied> τῇ Ἔπα
    <lb n="5" type="inWord"/><supplied reason="lost">γαθιαν</supplied>ῆ
      <expan>διώ<ex>ρυγι</ex></expan> <expan>Βακχιά<ex>δος</ex></expan>
    <lb n="6"/><gap reason="lost" extent="unknown" unit="character"/>
      <expan>Πατκ<ex>όννεως</ex></expan> τοῦ θεαγένους
    <lb n="7"/><gap reason="lost" quantity="6" unit="character"
      precision="low"/> <expan>μη<ex>τρὸς</ex></expan> Ταύρεως
    <lb n="8"/><gap reason="lost" extent="unknown" unit="character"/>
      <handShift new="m2"/><expan>σεση<ex>μείωμαι</ex></expan>
  </ab>
</div>
```

Corresponding to the typical Leiden print transcription:

```
[ἔτους α (?) Αὐτοκράτορος] . . [ ] . . του
[- ca.12 -] Σεβαστοῦ
[εἴργ(ασται) ὑ(πὲρ) χω(ματικῶν) ἔ]ργ(ων) τοῦ αὐτοῦ πρώτου (ἔτους)
[ -ca.?- ] κ κς ἐ[ν] τῇ Ἔπα -
[γαθιαν]ῆ διώ(ρυγι) Βακχιά(δος)
[ -ca.?- ] Πατκ(όννεως) τοῦ Θεαγένους
[ . . . . . ] μη(τρὸς) Ταύρεως
[ -ca.?- ] (hand 2) σεση(μείωμαι)
```

While transforming the XML to Leiden+ through our XSugar grammar yields:

```
1. [ἔτους] [<#α=1#> (?)] [Αὐτοκράτορος] .2[.1].2του
2. [ca.12] Σεβαστοῦ
3. [(εἴργ(ασται)) (ὑ(πὲρ) χω(ματικῶν))] ([ἔ]ργ(ων)) τοῦ αὐτοῦ πρώτου ((ἔτους))
4. [.?] <#κ=20#>  <#κς=26#> ἐ[ν] τῇ Ἔπα
5.- [γαθιαν]ῆ (διώ(ρυγι)) (Βακχιά(δος))
6. [.?] (Πατκ(όννεως)) τοῦ Θεαγένους
7. [ca.6] (μη(τρὸς)) Ταύρεως
8. [.?] $m2 (σεση(μείωμαι))
```

As you can see, things such as the Greek letter "κ" on line four being the number "20" are implicit in print, but explicit in both EpiDoc and Leiden+. The user can work on either the Leiden+ or XML representation of the text, and we store only the XML representation in our data repository (that is, the Leiden+ representation is only an intermediate form used for editing and is transformed back to XML when saved). A traditional Leiden "print preview" is possible by applying the standard EpiDoc XSLT stylesheets to this XML. In theory, this particular XSugar grammar could be re-used by other EpiDoc projects wishing to enable the same kind of alternative markup.

1.3. Repurposable Design

Due to institutional requirements, the DDbDP and HGV datasets needed separate editorial control and publishing mechanisms. In addition, their control over different types of content necessitated different editing mechanisms for each component. These requirements informed the design of how SoSOL interacts with data under its control and how this design is repurposable for use in other projects. The two high-level abstractions of data made by SoSOL are "publications" and "identifiers" . Identifiers are unique strings which can be mapped to a specific file path in the repository, while publications are arbitrary aggregations of identifiers. By defining an identifier superclass which defines common functionality for interacting with the data repository, we can then subclass this to provide functionality specific to a given category of data. The SoSOL implementation for IDP2, for example, provides identifier subclasses for DDbDP transcriptions, HGV metadata, and HGV translations. Editorial boards consequently have editorial control for only certain subclasses of identifiers. Publications in turn allow representation and aggregation of the complex many-

to-many relationships these components can have (for example, a document with two sides that may have one transcription and two metadata components). Packaging these related elements together both allows the user to switch between them and editorial boards to check related data which they may not have editorial control over but still require to make informed decisions about validity and approval. SoSOL can thus be integrated into other systems by implementing the identifier subclasses necessary for the given dataset as well as coherent means for aggregating these components into publications. One can imagine the simplest implementation as being an identifier whose name is the file path, which just presents the plaintext contents of the file for editing, and which has no relationships with other identifiers so that each publication is a single identifier.

1.4. Conclusions

Though SoSOL is still under development, early reception and feedback at training workshops conducted to introduce users to the system has been good. Thousands of texts have been created, edited, or corrected, and submitted through the editorial boards for voting, peer review, and publication. In addition to being publicly viewable through the Papyrological Navigator, these changes are regularly mirrored to a public copy of the Git data repository available on GitHub, where anyone may view and download the complete revision history (see previous footnote). Under IDP3, the currently-active final phase of the IDP project, user experience studies are being conducted and feedback from this process will be incorporated into the system.

In addition to the possible reuse of the system itself, we hope the more general goal of making data available with the full revision history will become a more widely adopted practice in digital humanities. Particularly for collaborative editing projects, this can allow anyone to see how changes have actually been effected, even outside of the specific editing environment employed. We feel that using a distributed version control system such as Git as the core data backend is conducive to this goal, as it enables easy distribution and updating of the data set alongside its complete version history. Thus, even if our particular editing environment is eventually outdated and replaced, or the data needs to be interacted with and edited using other mechanisms, the data backend can still be used for the next generation of tools and scholars.

References

Ryan Baumann. *IDP Data available on GitHub.* `http://digitalpapyrology.blogspot.com/2011/01/idp-data-available-on-github.html` (accessed 3 March, 2011).

Notes

1. We regard SoSOL as the intellectual heir of the Suda Online project, available: `http://www.stoa.org/sol/`.
2. JRuby, available: `http://www.jruby.org/`.
3. Ruby on Rails, available: `http://rubyonrails.org/`.
4. XSugar provides a means of converting between equivalent XML and non-XML vocabularies. See XSugar, available: `http://www.brics.dk/xsugar/`.
5. Git, available: `http://git-scm.com/`.
6. See Baumann (2011) (`#baum2011`) .

New Models of Digital Materialities

Blanchette, Jean-François
blanchette@gseis.ucla.edu
Information Studies, UCLA

Drucker, Johanna
drucker@gseis.ucla.edu
Information Studies, UCLA

Kirschenbaum, Matthew
mgk@umd.edu
Department of English; Maryland Institute for Technology in the Humanities, University of Maryland

One persistent myth of the digital age is that it differs fundamentally from all previous information epochs because in digital form information has finally achieved the long-standing historical aspiration to unburden itself from the shackles of matter. As a mere collection of 0s and 1s, digital information is imagined to be independent of the particular media on which it is stored—hard drive, optical disk, etc.—and the particular signal carrier which encode bits, whether magnetic polarities, voltage intensities, or pulses of light. Digital information also achieves a separation of content and form that could only be partially realized with analog carriers. This fantasy has implications for the ways we think about design, preservation, storage, use, and every other aspect of digital media. What can the digital humanities learn from and contribute to an engagement with the many aspects of the materialities of information?

The authors of these three papers undertake a common questioning of this purported independence from matter, a concept that has two distinct and important consequences: (a) the idea that digital information can be reproduced and distributed at negligible cost and high speed, and thus, is immune to the economics and logistics of analog media; (b) and that it can be accessed, used, or reproduced without the noise, corruption, and degradation that necessarily results from the handling of material carriers of information. Thus the concept of digital information as immaterial is fundamental to the ability of the digital to upend the analog world, and the foundation of a belief that any media that can be digitized or produced digitally will eventually succumb to the logics of digital information and its circulation through electronic networks—an argument powerfully encapsulated by Negroponte's slogan, "from atom to bits."

Such widespread assumptions have obscured the specific material constraints that obtain in digital environments. Only recently that the issue has emerged as a legitimate concern for scholarly enquiry —engaging concepts of materiality from literary, visual, media, and cultural studies and bringing them to bear on the analysis of digital environments. The papers in this session take up some of these issues by reading the machines, the specific properties of digital media from surface screen to deeper structures, as a demonstration of the ways the materiality of digital media can be engaged. The purpose of this work is to inform some of the basic tasks of digital humanists – the interpretation of digital media artifacts, the skills sets necessary to interrogate these artifacts, but also, our responsibility for the preservation and use of these objects as part of our cultural legacy.

PAPER 1

"Infrastructural Thinking" as Core Computing Skill

Blanchette, Jean-François
blanchette@gseis.ucla.edu
Information Studies, UCLA

It is often suggested that all digital humanists would benefit from learning programming. Through the acquisition of this core skill, they would engage with the practice that defines computing and directly experience its possibilities and constraints. Beyond mere mastery of a language, programming would expose them to formal methods for abstracting and modeling concepts and real-world phenomena. The current wave of interest in a "computational thinking" pedagogical paradigm mirrors this argument: computer science is primarily about modeling and abstraction of phenomena in ways amenable to algorithmic processing.

In this paper, I argue that programming, or its more complex formulation, "computational thinking," provides only a partial picture of computing, and correspondingly, only a partial skill set. A fuller picture requires engagement with the material foundations of computing. I use "material" here in a very literal

sense, to point to the physicality of bits (their encoding as magnetic polarities, voltages, etc) and the material constraints of the devices that process, store, and transport them. While the material dimension of computing constantly informs the practices of the computing professions, this dimension is also repressed, in the context of a general discourse that has emphasized the abstract dimension of the digital over its material substrate. Yet, this materiality, perhaps unexpectedly, holds the key to analyzing the shape and evolution of the computing infrastructure. And while digital humanists may well benefit from engaging in "computational thinking," I will argue the computing infrastructure implicitly performs much of that thinking, before a single line of application code is written.

While programming deals with creating applications that provide service to users, infrastructure software provides services to applications, by mediating their access to computing resources, the physical devices that provide processing power, storage, and networking. Infrastructure software is most commonly encountered in the form of operating systems, but is also embedded in hardware (the firmware in a hard drive) or in specialized computers (e.g., web servers, or routers). Whatever its specific form, the role of infrastructure software is to provide a series of transformations whereby the signals that encode bits on some physical media (fiber optic, magnetic drive, electrical wires) become accessible for symbolic manipulation by applications. Infrastructure software must be able to accommodate growth in size and traffic, technical evolution and decay, diversity of implementations, integration of new services to answer unanticipated needs, and emergent behaviors, among other things. It must provide programmers with stable interfaces to system resources in the face of continuously evolving computing hardware—processors, storage devices, networking technologies, etc.

The computing industry manages to accomplish this feat through the design strategy of *modularity*, whereby a module's implementation can be designed and revised without knowledge of other modules' implementation. Modularity performs this magic by decoupling functional specification from implementation: operating systems, for example, enable applications to open, write to, and delete files, without any knowledge of the specific storage devices on which these files reside. This decoupling provides the required freedom and flexibility for the management, coordination, and evolution of complex technical systems. However, in abstracting from specific implementations of physical resources, such decoupling necessarily involves efficiency trade-offs. The TCP/IP protocols for example provide abstractions of networks that favor resilience (the network can survive nuclear attacks) over quality of service (the network provides no minimum delays for delivery of packets). Applications sensitive to such delays (e.g., IP telephony or streaming media) must thus overcome the infrastructural bias of the protocols to secure the quality of service they require.

An important point is that efficiency trade-offs (or biases) embedded in a given modular organization become entrenched through their institutionalization in a variety of domains: standards, material infrastructure (e.g., routers), and social practices (e.g. technical training) may all provide for the endurance of particular sets of abstraction. This entrenchment is further enabled by the economies of scale such institutionalization affords. An immediate consequence is that the computing infrastructure, like all other infrastructures, is fundamentally conservative in character. Yet, it is also constantly under pressure from the need to integrate changes in the material basis of computing: multi-core, cloud-based, and mobile computing are three emerging material changes that will register at almost every level of the infrastructure.

Computing, it turns out, is material through and through. But this materiality is diffuse, parceled out and distributed throughout the entire computing ecosystem. It is always in subtle flux, structured by the persistence of modular decomposition, yet pressured to evolve as new materials, requiring new tradeoffs emerge. This paper thus argues that, in a very literal and fundamental sense, materiality is a key entry point for reading infrastructural change, for identifying opportunities for innovation that leverage such change, and for acquiring a deep understanding of the possibilities and constraints of computing. This understanding is not particularly provided by exposure to programming languages. Rather, it requires familiarity with the conflicts and compromises of standardization, with the principles of modularity and layering, and with a material history of computing that largely remains to be written.

PAPER 2

Performative Materiality and Interpretative Interface

Drucker, Johanna
drucker@gseis.ucla.edu
Information Studies, UCLA

Approaches to interface design have come mainly from the HCI community, with an emphasis on maximum efficiency in the user-centered experience. Since the days of Douglas Engelbart and Ivan Sutherland's experiments with head sets, pedals, mice, and screens, in work that led to the development of the Graphical User Interface, the dominant paradigm in the human-machine relationship has come from an engineering sensibility. Leading practitioners in that field, from Stuart Card, Ben Shneiderman, and others, have defined basic principles for design methodology and display that are premised on a mechanistic analysis of user's abilities to process information effectively. This approach, taken from flight simulators and applied to the vast numbers of tasks for searching navigating, buying, and communicating online, is grounded in a user-as-consumer model. Criticisms from inside that community, such as the work of Jesse James Garrett (showing the confusion between information and task based approaches) or Aaron Marcus's group (analyzing cultural differences and their connection to interface functionality) have provided useful insights and shifted design principles to be more nuanced. But the basic model of the user-centered approach to interface design remains in place. And it has been adopted by humanists, particularly when the resources to do so are available.

If we bring the legacy of critical theory to bear on this model, however, we see that the same critique leveled by post-structuralists against New Criticism is pertinent here. The "text" of an interface is not a thing, stable and self-evident, whose meaning can be fixed through a detailed reading of its elements. An interface is a site of provocation for reading, and, in the same manner as a film, literary work, or any other "text" (fashion magazine, instruction manual), it is a space for interpretation involved an individual *subject*, not a generic user. In critical parlance, both an enunciating and enunciated subject – the speaking and the spoken subject – are aspects of textual production. (Text here is meant broadly.) This concept of performativity, articulated by John Austin in *How to do things with Words*, has echoes within the field of anthropology, gender studies, and cultural studies. By situating texts and speakers within pragmatic circumstances of use, ritual, exchange, and communities of practice, performativity stripped away any foundation for thinking meaning was inherent in a text or work. Performativity offered a sharp rebuke to notions of agency (individuals) and autonomy (of texts).

How can we, that is, the community of digital humanists, take these critical insights from literary, cultural, and gender studies into our current practice? If the object is merely to demonstrate that one may read an interface with the same techniques we used to read *Young Mr. Lincoln* or to follow Laura Mulvey's arguments into a new realm of semiotic analysis, a rather tedious and predictable path would like ahead. This might have some value in the undergraduate classroom, as the unpacking of ideological subtexts fascinates the young. But for those of us concerned with the design of environments for digital humanities and its research agendas, the questions that arise from this critical encounter are quite different. Can we conceive of models of interface that are genuine instruments for research? That are not merely queries within pre-set data that search and sort according to an immutable agenda? How can we imagine an interface that allows content modeling, intellectual argument, rhetorical engagement? In such an approach, the formal, graphical materiality of the interface might register the performative dimensions as well as support them. Such approaches would be distinct from those in the HCI community in terms of their fundamental values. In place of transparency and clarity, they would foreground ambiguity and uncertainty, unresolvable multiplicities in place of singularities and certainties. Sustained interpretative engagement, not efficient completion of tasks, would be the desired outcome.

This is not an argument in favor of bad design. Nor is it a perverse justification for the ways in which under-resourced projects create confusion, as if that were a value for humanists. Quite the contrary. The challenge of creating an interface in which the performative character of interpretation can be supported and registered builds on demonstrable principles: multiple points of view, correlatable displays, aggregated data, social mediation and networking as a feature of scholarly work, and some of the old, but still charming, qualities of games like Nomic, with their emerging rule sets.

My argument is that the humanities embody a set of values and approaches to knowledge as interpretation that cannot be supported by a mechanistic approach

to design. This is not just a semantic exercise, but a point of departure for implementation. The concept of performative materiality has a double meaning here. In the first sense, materiality is understood to produce meaning as a performance, just as any other "text" is constituted through a reading. That notion is fundamental to humanistic approaches to interpretation as situated, partial, non-repeatable. In the second sense, performative materiality suggests an approach to design in which use registers in the substrate and structure so that the content model and its expressions evolve. The "structure of knowledge" becomes a "scheme of knowing" that inscribes use as well as provoking it. The idea of a user-consumer is replaced by a maker-producer, a performer, whose performance changes the game. This takes us back to some of the earlier theory of games, to the work of Brenda Laurel and others, whose theoretical training brought notions of subjectivity and performance into the study of online environments.

This paper does not claim to have a toolset of design solutions, since by definition, that would put us right back into the HCI model. Instead, it is an attempt to lay out some basic ideas on which to imagine a performative approach to materiality and the design of an interpretative interface. Such an interface supports acts of interpretation (does not merely return selected results from a pre-existing data set) and also is changed by acts of interpretation, it evolves. Performative materality and interpretative interface are co-dependent and emergent.

PAPER 3

Checksums: Digital Materiality in the Archive

Kirschenbaum, Matthew
mgk@umd.edu
Department of English; Maryland Institute for Technology in the Humanities, University of Maryland

---

The general conversation about "materiality" in digital media has been ongoing for quite some time (notably Markley 1997; Hayles 2002). A number of new foci, models, and constructions have also recently been introduced into the conversation around the term. These include Kirschenbaum's "formal materiality" and accompanying work on digital forensics (2008), the "media archeology" paradigm emerging out of several key European writers following in the wake of Friedrich Kittler (Parikka 2007, Ernst 2005), and the platform studies approach developed and encouraged by Montfort and Bogost (2009). At the same time, in the archival community, practitioners are finding themselves confronted with the materiality of born-digital objects in palpable and often increasingly time-sensitive real-world ways: as more and more collections begin to process and receive digital storage media as elements in the acquisition of personal "papers" from writers, politicians, and other public figures, those charged with their long-term care are implementing their own working models of materiality as they make decisions about what to save, what to index, and what to provide access to. (Given the venue for this year's conference, it's worth noting that Stanford University Libraries has been a pioneer in this area, with such efforts as the Self-Archiving Legacy Toolkit (SALT), the AIMS project on Born-Digital Collections and New Models for Inter-Institutional Model for Stewardship, and their participation in the Preserving Virtual Worlds project.) This paper will therefore seek to evaluate recent developments in the theoretical conversation about digital materiality in the specific context of applied practice in the archival community. However, it will not assume that born-digital archival content can function only as a test-bed for the various theoretical models; instead, it will also look at how the decisions being made in archival settings have the potential to inform critical and theoretical discourse.

The paper follows a case-study approach. Thus, it will consider media archeology, which one definition glosses as "histories of suppressed, neglected, and forgotten media [. . .] ones that do not point selectively and teleologically to the present cultural situation and currently dominant media as their 'perfection'" (Huhtamo 2010) in light of the Deena Larsen Collection at the University of Maryland. Larsen, who is traditionally associated with the Eastgate stable of hypertext authors through work such as *Marble Springs* (1993) and *Samplers* (1997) has deposited a large and heterogeneous array of hardware, software, storage media (some eight hundred 3½-inch diskettes), notebooks, manuscripts, correspondence, and ephemera at the Maryland Institute for Technology in the Humanities. This material includes a number of items related to her best-known work, *Marble Springs* (1993), written in Hypercard and published the same year, it turns out, as a much more famous work of digital storycraft, Cyan's *Myst*, also authored in Hypercard. The Maryland collection includes the original text, with annotations, as a manuscript in a notebook, various electronic

drafts and early implementations, installed copies of the work (which allows the user to add marginal notes, thus creating the capacity to render every copy unique), and most unusually, a shower curtain which contains laminated screenshots of different nodes (lexia) pasted up and linked together with colored string to diagram the affective relations between the different elements of the text. How can media archeology, which, following Foucault's venerable formulations, seeks to tease out "hitherto unnoticed continuities and ruptures" (Huhtamo 2010) inform curatorial practice around this work? Among other things, this paper will argue for a network-oriented model of access, whereby a user of the collection, through metadata packaging, is encouraged to access and evaluate items in relation to one another rather than in isolation, as atomized treasures sprung from a Hollinger box (or its digital equivalent, a FEDORA record).

Similarly, the platform studies approach advocated by Montfort and Bogost will be evaluated in relation to a second recent project from MITH, a site devoted to vintage computing which uses a considered metadata and modeling approach to computing hardware, whereby individual components of the vintage machines are documented, contextualized within their relation to the system as a whole, and expressed using Dublin Core. While not "platform studies" in its own right, the extent to which formalized representations of vintage computing systems can serve as the basis for further work in that vein is important, since there are currently few precedents for cataloging actual artifacts of computer history, and in particular considering specific components as individualized entities rather than as generic classes of mass-produced material. The paper will therefore ask what kind of documentation and metadata a scholar interested in the affordances and individuality of a particular computing system would need in order to use a cataloged instance as the basis for critical work in the platform studies model.

Finally, the paper will consider Kirschenbaum's interwoven strands of formal and forensic materiality in relation to an event which post-dated his 2008 book, the recovery, reconstitution, and successful emulation of the original software program for William Gibson's famously self-effacing poem "Agrippa" (Kirschenbaum, et al. 2009). Emulation in particular allows us to usefully consider the armature of "formal materiality," since emulation functions precisely to naturalize or operationalize one baseline computing system as the virtual host for a working instantiation of another. Yet emulation is also an uncanny event, a kind of literalization of the age-old conceit of the "ghost in the machine" as a contemporary operating system becomes the proxy for one long (un)dead. "Agrippa" itself, with its themes of memory, media, and loss is an apt vehicle for a meditation on formal materiality and the limits of absolute emulation, especially since there is one overridingly obvious fact about any virtual implementation or emulation of "Agrippa": unlike the original, which, famously, was a one-off "run" encrypted with a military-grade one-way key, here one can spawn "images" of the original disk at will and pass them through the emulator time and again. Emulation, the paper concludes, is finally chimerical as is, perhaps, the notion of formal materiality itself: virtualization will always be interrupted by the individuality of original events.

One last word: the paper has a broader agenda beyond putting theory through its paces, using archival content as a "checksum" for critical debate. Digital humanities itself, as a community, has much to offer to ongoing efforts in the archives world, yet thus far there has been a curious gulf between the two, with only a handful of individuals serving as ambassadors between the two communities. If digital humanities can offer a forum in which the artifacts and objects of contemporary cultural heritage, many of which will be *born-digital* rather than digitized, can serve as the basis for critical and technical inquiry then it will be well positioned to take part in the increasingly urgent societal conversation around the future of our digital and material present.

The Theory and Design of PlotVis

Dobson, Teresa M.
teresa.dobson@ubc.ca
University of British Columbia, Canada

Ruecker, Stan
sruecker@ualberta.ca
The University of Alberta

Brown, Monica
mm2brown@interchange.ubc.ca
University of British Columbia, Canada

Rodriguez, Omar
omar.rodrigueza@gmail.com
The University of Alberta

Michura, Piotr
zemichur@cyf-kr.edu.pl
Academy of Fine Arts, Krakow, Poland

Grue, Dustin
dgrue@shaw.ca
University of British Columbia, Canada

PlotVis is a 3D system for people interested in examining narrative action, or plot, from a variety of perspectives. The system was developed as one outcome of a funded study examining the teaching of complex narrative forms in secondary and undergraduate classrooms. Results of this study have revealed, inter alia, that conventional approaches to teaching fiction, particularly at the secondary level, fail to take account of the diversity of contemporary narrative (cf Dobson, 2006).

For example, instructors, particularly at the secondary level, still rely heavily on the five-stage plot mapping first described by Gustav Freytag (1863) in Die Technik des Dramas (Figure 1), and the superimposition of this model on forms beyond those it was originally intended to describe can be misleading (Dobson, Michura, Ruecker, 2010).

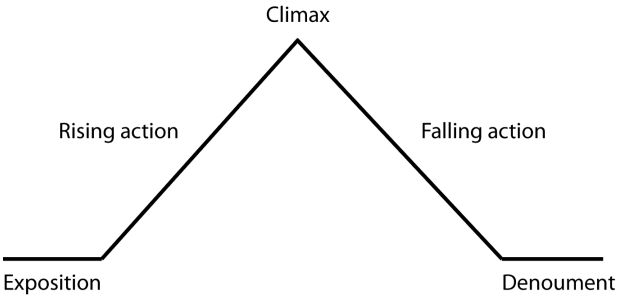


Figure 1: Freytag's Pyramid shows five basic components of plot, based on Greek and Shakespearean tragedy.

In addition, the reliance on the Cartesian graph to model narrative generally – and there are a number of examples of such graphing (e.g., Sterne, 1847; Vonnegut, 1973) – is restricting because it fails to take account of the multiple dimensions of story.

Considering the limitations of the Cartesian graph for modeling narrative along with conceptualizations of narrative as multidimensional (e.g., Shields, 2000), our goal was to produce a three-dimensional digital environment that reifies different perspectives on plot, so that students and other scholars can quickly shift from one 3D object to another, or spend time exploring any of the visualizations in more detail. We currently support three perspectives and are beginning to experiment with the design of a fourth.

This panel will examine the results of the PlotVis project from several different perspectives, including an introduction to the project, schema design for narrative encoding, a reading practices study, and design and programming.

2. Introduction to PlotVis as a Form of Distant Reading

" … what we really need is a little pact with the devil: we know how to read texts, now let's learn how not to read them. Distant reading: where distance, let me repeat it, is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems. (Moretti, 2000, p. 57)."

Recently, literary theorists and digital humanists alike have taken up Moretti's appeal, later developed in *Graphs, Maps, Trees* (2005), for the development of new methods of textual analysis. Clement (2008), for example, uses mapping techniques and visualizations to facilitate, quite literally, new perspectives on a canonical text, Gertrude Stein's *The Making of Americans*. By premising knowledge on distance rather than closeness, Clement discerns significant patterns in the narrative's structure, in so doing

discovering a new logic to Stein's text. Such new insights, made possible by adapting the scale and focus of textual analysis, are exactly what Moretti expects readers will earn by learning how not to read texts—and, by extension, by learning how to use different technologies to complement, and corroborate, our readings of texts.

Experiments with distant reading challenge the truism that readers make sense of a story by seizing upon a narrative episode, using it as a key to gain entry into the deeper meaning inherent within the text, "as if meaning resided in a buried treasure chest or behind a lock door" (Clement, 2008, p. 365). Clement proposes a different metaphor for understanding narrative structure, that of a map key, which instead implies that readers make sense of stories by using narrative structure as a guide. The metaphor of the map key offers a view of reading complex narrative as a process of orientation, through which readers manage their awareness of a text's meanings by acquainting themselves with its structure.

Similar views of distant reading and complex narrative inform the project of developing XML schemas for encoding digitized versions of fictional narratives. Undertaken as part of a larger study based out of the University of British Columbia, this project investigates, and proposes, methods of reading adapted to contemporary print and digital fiction's "shift away from conventional narrative logic toward indeterminacy, fragmentation, and open-endedness" (Dobson, 2008, p. 1). While the goal of our larger study is the development of new models for writing, reading, and teaching complex print and digital narrative, the focus here will be the use of XML to encode fictional narratives as, in itself, a form of textual analysis, involving both closeness and distance, and enabling even further alterations, of the kind Moretti proposes, to the scale and focus of scholarly reading practices.

3. Towards a Schema Design for Narrative Encoding

This paper will discuss the iterative development and use of a set of Extensible Markup Language (XML) schemas for encoding digitized versions of fictional narratives. Although the Text Encoding Initiative (TEI) Guidelines for Electronic Text Encoding and Interchange define and document a thorough markup language for encoding humanities texts, including significant provisions for encoding scholarly editions of drama, poetry, and manuscripts, there presently exist only a handful of XML schemas designed specifically for marking up literary fiction (e.g., StoryML, FicML, PftML).

Drawing on principles and approaches from narratology, our literary encoding schemas combine TEI elements and attributes with tags that specifically mark up elements of narrative structure, such as actions, characters, dialogue, narration, objects, places, and time. Examples of encoded short stories, such as Alice Munro's "The Love of a Good Woman" and Ernest Hemingway's "Hills Like White Elephants," will be used to demonstrate, and critique, the different possibilities we have explored for developing literary XML schemas.

4. From Envisaging to Visualization: Young People's Narrative Reading Practices

This paper will present findings of a study examining the reading practices of young adults. The study focused on the dialectic between how students of literature *envisage* narrative and how they actually visualize it in systematized approaches. Fifty participants in grades 11 and 12 read Hemingway's "Hills Like White Elephants" and O'Faolain's "The Trout." They then engaged in a series of activities: evaluating form by physically "cutting and pasting" a narrative's text, sketching narrative diagrams, XML tagging of narratives, and an exit interview.

Participants indicated that narratives were enjoyable because their plots were "unexpected," but also un-enjoyable because they were "hard to understand." That is, comprehension of a narrative is a condition of enjoyment and disappointment: narratives ought to be original but not confusing. In XML tagging, grade 11 students were given a basic schema to tag a narrative but were also encouraged to modify this schema. The schemas were altered beyond a systematized approach, instead revealing critical activity across a range of "levels." Globally, for example, subjects modified the schemas beyond a "systematic" approach in order to highlight salient features particular to a text. Subjects added the tags *setting* and *contradictory statements* to the XML schema for "Hills," suggesting a post-critical (and almost juridical) attitude. More locally, there is evidence that tagging is dependent on "spheres of lexical activity," where tagging of features depends on the immediate context in which they occur. Furthermore, tagging appears to be not only independent of syntax, but also independent of the tagged-narrative's grammar – subjects prescribe systematic rules, but do not follow them.

5. PlotVis: Design and Implementation of Five Plot Models

Our final panelists will discuss issues related to designing and programming four different views of narrative structure. In the first case, the user is able to see a Fibonacci series where one of the encoded plot elements is placed at the centre of the visualization, and other elements are arranged out from the centre, either sequentially as they appear in the text or else hierarchically as they are arranged in the XML tree (Figure 2).

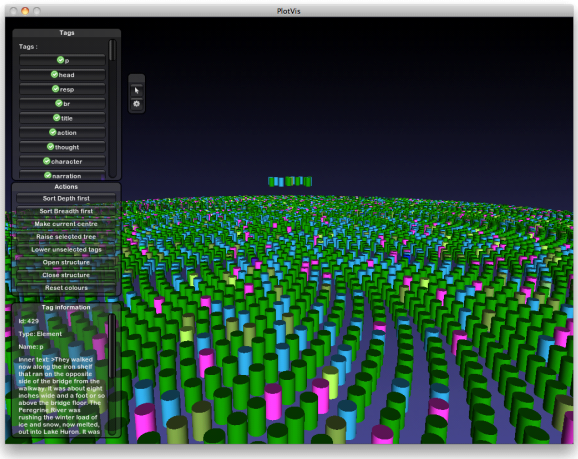


Figure 2: This screenshot of Munro's "Love of a Good Woman" demonstrates the complexity of the tagging, with each coloured cylinder representing one of roughly 7000 tagged pieces of text.

In this visualization, each piece of text appears as a coloured cylinder. The height of the cylinders can be associated with the length of the text they contain. The second visualization privileges sequence over centrality, with multiple pipelines representing the characters or narrator, and changes to the arrangement of the pipe showing where changes in time occur (Figure 3).

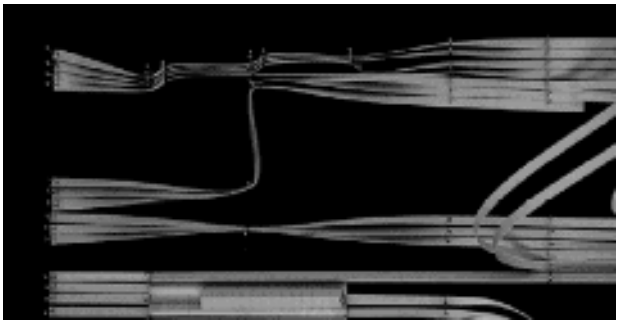


Figure 3: Our design sketch for a view that privileges sequence over centrality, with multiple pipelines representing the characters or narrator.

The third visualization is based on an architectural metaphor, where readers see the different kinds of encoded plot elements as floors of a building (Figure 4).

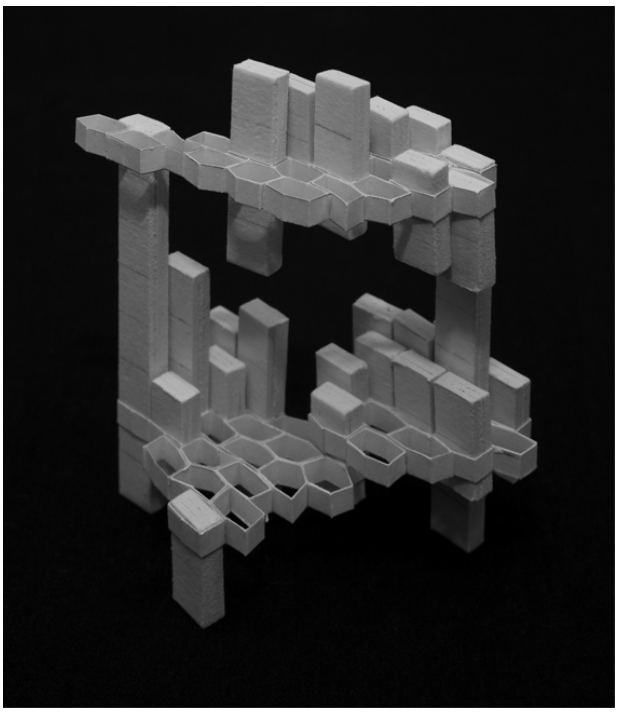


Figure 4: Our design sketch of the architectural concept of story visualization.

Text that occurs only within a single element tag fills a space on a single floor, while text that spans multiple elements appears as blocks that span the floors. Our fourth and most recent design resembles in some ways our sequential model, only in this case the text is represented by vertical walls of text that bend at locations where there are changes in time, while the other encoded plot elements appear as colours on the surfaces of the walls (Figure 5).

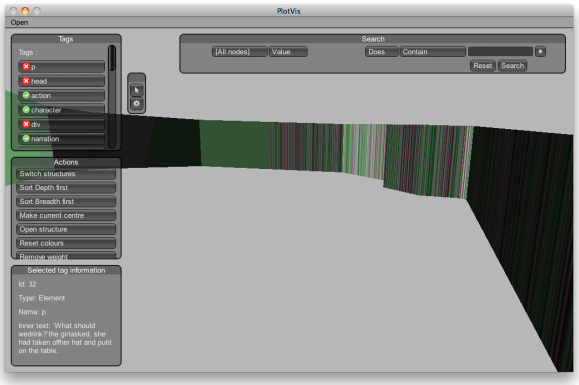


Figure 5: This screenshot shows the wall display. Vertical coloured bands represent the encoded text elements that are shown as nodes in Figure 2.

The visualization system was created using a popular game engine, called Unity3D, which provides us with a flexible platform for interaction with 3D objects, primarily through writing Javascript. The contents of the visualization, as indicated above, are produced from XML-encoded text files.

References

Clement, T. E., Dobson, T.M. (2008) (2008). ''A thing not beginning and not ending': using digital tools to distant-read gertrude stein's the making of americans'. *Literary and Linguistic Computing. Reading, Writing, and Teaching Complex Narrative. Standard Research Grant Application submitted to the Social Sciences and Humanities Research Council of Canada.*, pp. 361-381 23(3): 361-381

Dobson, T. M. (2006). 'For the love of a good narrative: Digitality and textuality'. *English Teaching: Practice and Critique.* , pp. 56-68.

Dobson, T. M., Michura, P., Ruecker, S. (2010). 'Visualizing plot in 3D'. *Proceedings of the Fourth International Conference on Digital Society*.

Freytag, G. (1863/1983). *Die technik des dramas*.

Hemingway, E. (1927). *Men Without Women*.

Jessop, M. (2008). 'Digital visualization as a scholarly activity'. *Literary and Linguistic Computing.* 23(3): 281-293.

McCarty, W. (2002). 'Humanities computing: essential problems, experimental practice'. *Literary and Linguistic Computing .* 17(1): 103-125.

Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary theory*.

Munro, A. (1998). *The love of a good woman*.

O'Faolain, S. (1980). 'The trout'. *Collected stories. (Vol. I).* Pp. 383-386.

Shields, C. (2000). '"Ilk"'. *Dressing up for the carnival.* Pp. 53-60.

Sterne, L. (1847). *The works of Laurence Sterne, containing the life and opinions of Tristram Shandy, gentleman*.

Van Peer, W., Chatman, S. B. (eds.) (2001). *New perspectives on narrative perspective*.

Vonnegut, K. (1973). *Breakfast of champions*.

Modeling Event-Based Historical Narratives: A Conversation Between Digital Humanists, Information Scientists and Computer Scientists

Meeks, Elijah
emeeks@stanford.edu
Stanford University

Mostern, Ruth
rmostern@ucmerced.edu
University of California, Merced

Grossner, Karl
karlg@geog.ucsb.edu
University of California, Santa Barbara

Shaw, Ryan
ryanshaw@unc.edu
University of North Carolina at Chapel Hill

Jain, Ramesh
jain@ics.uci.edu
University of California, Irvine

Kantabutra, Vitit
vkantabu@coe.isu.edu
Idaho State University

As digital humanities scholars increase the technological sophistication of their methodologies and tool use to represent historical change, their work becomes more theoretically complex from an information and computer science perspective. While much DH scholarship can be performed using off-the-shelf tools and simple data models, attempts to model highly nuanced historical knowledge, especially in event-based representation, requires a greater theoretical investment. This panel will present humanities scholars whose work focuses on using graph and semi-graph data models that track historical change as events in different parts of the digital scholarly media ecosystem—from the theoretical layout of the data and the representation of historical fact to the creation and population of a data structure to the analysis, representation and interaction with the product—in a dialectic with computer scientists and information scientists who have grappled with the problems of designing robust

models and implementing performant technology in this domain. Through analysis of the digital scholarly media process from theory to creation to review, we hope to reveal sophisticated practical and theoretical solutions for representing, creating and reviewing historical knowledge.

The panel will be chaired by Elijah Meeks and allow for a short presentation by each of the discussants on the topics described below. This will be followed by a roundtable-style discussion with media examples from each discussant's work cued up so that, if the conversation focuses on one of these works, it can then be demonstrated and examined in detail. The guiding themes of the panel are theoretical challenges of event-based historical models, practical issues in creating sophisticated digital scholarly media using such models, and methods to review the models and the media created from them.

The panel presents a continuum of scholars from the humanities through information science to computer science, with Elijah Meeks (Digital Humanities Specialist, Stanford University) and Ruth Mostern (Associate Professor of History, University of California, Merced) representing the traditional humanities scholars. Elijah Meeks will speak about his work transitioning the *Mapping the Republic of Letters* database into a pure graph model and the challenges and benefits that this representation of historical knowledge presents in contrast to traditional relational databases as well as the suitability of graph data for event modeling. Ruth Mostern will focus on the role of event modeling in her current Yellow River research, tracking the linkages between historical environmental change and socio-political change, with an emphasis on the temporal and spatial reconciliation of differing scales and emphases of data, specifically comparing historical data collected for environmental science and traditional textual sources for tracking political and social change over time and space.

Karl Grossner (Geography, University of California, Santa Barbara) and Ryan Shaw (Assistant Professor of Information and Library Science, University of North Carolina at Chapel Hill) are both information scientists whose work has focused on issues critical to historical narrative modeling: ontologies and periodization. Ryan Shaw will speak to the complexities of event representation, focusing on periods and events as not only existing in the past, but also produced by discourse about the past, and the need for an approach to representation in which we can move back and forth between treating periods and events as, on the one hand, shared intersubjective points of reference and, on the other, unique perspectives on the past articulated through narrative techniques. Karl Grossner will discuss his work developing a spatial history ontology in which historical processes are modeled as theories of event relations, as well as its implementation in a PostgreSQL spatial database able to support mapping, analytical applications and representations of narrative.

Finally, Ramesh Jain (University of California, Irvine) and Vitit Kantabutra (Idaho State University) will discuss the practical issues of implementing such theoretical constructs in software. Vitit Kantabutra will spend some time explaining his work with Jack Owens fleshing out the Intentionally Linked Entity model, which is a generalized graph data model with a wide range of applications, and is especially suitable for the representation of historical and spatiotemporal events. Ramesh Jain will engage with the processing of event-based data semantically, both through typical searches and by exploring the emergent attributes of the database as an object itself.

By presenting a continuum of academic specialties that are unified in their focus on sophisticated digital representation of historical knowledge, we feel this panel will provide useful practical and theoretical value to the Digital Humanities community. We also hope to add to the growing discussion of how to judge the quality of the media produced in these endeavors, not only from the perspective of humanist scholars but from scholars in the field of information science and computer science.

Networks, Literature, Culture

Moretti, Franco
moretti@stanford.edu
Stanford University

Finn, Ed
edfinn@stanford.edu
Stanford University

Lewis, Rhiannon
rmlewis@stanford.edu
Stanford University

Frank, Zephyr
zfrank@stanford.edu
Stanford University

Overview: Our panel proposal, and the three individual papers, present work in progress at the Stanford Literary Lab, a work space opened in the Fall of 2010 where faculty and students conduct research of a digital and quantitative nature. Ideally, we try to generate critical "experiments" that mix archival exploration and hypothesis testing, and extend over a period of one-two years; all of them collaborative, and developing through regular group meetings that evaluate results and plan the work to come.

PAPER 1

Reading, Writing and Reputation: Literary Networks in Contemporary American Fiction

Finn, Ed
edfinn@stanford.edu
Stanford University

1. Overview

Long-established models of literary production are changing dramatically as the digital era continues to blur, and at times erase, the divisions between authors, critics and readers. Millions of cultural consumers are now empowered to participate in previously closed literary conversations and to express forms of mass distinction through their purchases and reviews. My project argues that these traces of popular reading choices constitute a fresh perspective on elusive audience reactions to literature, one that reveals distinct networks of conversation that are transforming previously well-understood relationships between writers and their readers, between the art of fiction and the market for books (Radway). Employing network analysis methodologies and 'distant reading' of book reviews, recommendations and other digital traces of cultural distinction (Moretti), my research develops new models for studying literary culture in America today. In this paper I consider the reception of three mid-career writers, David Foster Wallace, Junot Díaz and Colson Whitehead, asking how they have redefined authorial expectations and literary identity through their work.

2. Background

My project is founded on the argument that as literary production evolves, new kinds of reading communities and collaborative cultural entities are emerging. Many of these communities are ephemeral and quite often they are fostered by commercial interests seeking to capitalize on their cultural production. Nevertheless, a handful of websites like Amazon continue to dominate the marketplace for books and attract millions of customer reviews, ratings and purchase decisions, and the literary ecologies of these book reviews have become valuable research resources. The ideational networks I explore are made up of books, authors, characters and other literary entities (these are the nodes), along with the references linking them together as collocations in book reviews, suggestions from recommendation engines, and other structures of connection. As my project has moved from Thomas Pynchon and Toni Morrison to the younger generation represented by Wallace, Díaz and Whitehead, I have improved my data-gathering and network analysis methodologies. With new data and better tools, I hope to determine how the game of authorial fame is changing in an increasingly reflexive, networked literary landscape.

3. Proposal

This paper will present my research on three younger writers who have broken new ground in literary constructions of identity in a shifting landscape of reception. Wallace, Díaz and Whitehead, all members of "Generation X," are writers who have captured national attention through particularities of self-presentation and novelistic style.

I argue that these authors signal a sea change in literary reputation. Authors as well as publishers are now interacting with active communities of readers who conduct complex cultural conversations independent of traditional arbiters of taste. As the barriers separating readers from ordained critics crumble online, younger authors are increasingly engaging with audiences that are both collaborative and vocal. Tracing the half-spun career arcs of Wallace, Díaz and Whitehead, this paper articulates a new model of contemporary literary culture: a reading society that demands increasing authorial reflexivity to mirror the collaborative, iterative nature of digital literary conversations.

Each author makes a distinct form of literary identity central to his work. David Foster Wallace defined a deeply introspective and reflexive narrative voice, pioneering a style so individual that he ultimately felt it had become clichéd and struggled to escape it in his final years of writing. In his breakout novel *The Brief, Wondrous Life of Oscar Wao*, Junot Díaz has tackled the challenges of his hybrid Dominican/American identity by creating a new argot of Spanglish phrases, pop cultural references and an ingenious deployment of footnotes to both buttress and undermine normative American understandings of Latin American history and culture. Finally, Colson Whitehead has similarly inverted expectations for African American authors, accomplishing the seemingly impossible feat of writing in the traditions of both Toni Morrison and Thomas Pynchon. As a group, these writers have little in common except their age, positions of cultural prestige and their talents as re-inventors of literary identity. This makes them excellent subjects for a study of contemporary literary reception and the collective construction of literary identities.

4. Methodology

The project draws on two primary datasets: first, a corpus of professional and consumer book reviews collected from nationally prestigious reviewing newspapers and magazines along with consumer reviews from Amazon dating back to 1996; second, networks of recommendations based on consumer purchases and book ownership drawn from the websites Amazon and LibraryThing. Comparing these different literary networks allows me to make broader arguments about contemporary authorial fame and the changing role of the everyday reader in literary conversations.

For the first dataset, consumer and professional book reviews, I have employed the MorphAdorner project's Named Entity Recognition tool to create a dictionary of literary proper nouns (authors, titles, character names, etc). This dictionary was used to identify collocations on a paragraph level throughout these reviews, making nouns into nodes and collocations into links. This approach allows me to empirically identify and visualize those authors and texts frequently mentioned in the same review contexts. The resulting noun networks reveal not only the distinctions between everyday readers and more traditional arbiters of literary taste but the ways in which popular authors are increasingly carrying on multiple independent and complex literary conversations.

The second dataset combines book recommendations provided by Amazon's "Customers who bought this also bought" engine and LibraryThing's recommendation engine. In this case books form the nodes and recommendations the links connecting them. Employing network analysis to identify patterns of prestige and clustering in these recommendation networks has allowed me to trace the diverse functions of genre and authorial identity in ordering literary texts and to identify the basic rules of contemporary market canonicity.

5. Conclusion

As I have moved this project through its first two case studies, with Pynchon and Morrison, I have refined a hybrid methodology that explores literary and cultural context through a limited empirical lens, considering particular digital and textual networks of reference and evaluation. This iteration of the process takes on three authors with very different styles to ask how a younger literary generation is working together with newly empowered readers to construct new kinds of multiply mediated, widely collaborative forms of cultural identity. By studying these writers mid-career, I hope to trace both the evolution of contemporary authorial fame and its relationship to wider systems of social distinction in a rapidly evolving digital landscape.

References

Bourdieu, P. (1993). *The Field of Cultural Production: Essays on Art and Literature..* Randal Johnson (ed.). 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.

(2009). *MorphAdorner..* Northwestern University. http://morphadorner.northwestern.edu/.

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.

(2010). *yEd Graph Editor.* yWorks GmbH. http://www.yworks.com/.

PAPER 2

Paper Two: Plot as Network: Quantifying the Evolution of Dramatic Style

Lewis, Rhiannon
rmlewis@stanford.edu
Stanford University

In the last decade or so, quantitative evidence has become part of literary study in a variety of ways – from the material realities of book history [library holdings, or translation flows], to the linguistic macro-patterns that have renewed the study of attribution, genre differentiation, and stylistics. Plot, however, has proved much more difficult to quantify. This paper looks for a possible solution in network theory, whose concepts re-define plot as a system where characters are the vertices, and their interactions the edges of a narrative network. Focusing, for now, on dramatic literature, such "network narratology" brings to light the striking discontinuities between the structure of ancient and Renaissance tragedy, and suggests a reconceptualization of literary characters in terms of their "connectedness" and their position within the network.

This paper grew out of Franco Moretti's network analysis of three Shakespeare plays: *Hamlet*, *Macbeth*, and *King Lear*. Preliminary work for the project included php-extraction of speaker-receiver data from MIT's xml-encoded Shakespeare corpus. The program *R* was then used to generate network analyses. A team of Stanford graduate students hand-corrected receiver attributions to a level of 100% accuracy. The paper's methodological component discusses the difficulties of automated recipient attribution in dramatic texts, and touches on issues of "quantifying" performance literature, including plays lacking an authoritative text. The idea of (performance) time turned into space is grounded in narrative theory: specifically, Alex Woloch's concept of a character-*system* made of many character-spaces (*The One vs. the Many*). In addition to modeling "plot as network" based in dialogue, I also utilize stage directions to map networks as shared "space." Here, space includes more ambiguous character presence such as disguised identity and plays-within-plays, as well as the auditory presence of eavesdropping. Thus, overlapping networks of character position -- both physical and verbal -- model multiple dimensions of plot through form and performance.

The paper draws on extant and original corpora of over 70 plays, from classical through Renaissance drama. They include Shakespeare's complete dramatic works, as well as select plays by Marlowe, Jonson, and Webster. The classical plays are drawn from Project Perseus' database. Evaluating our script's performance across literary periods has revealed that classical dramatic dialogue follows a linear progression more frequently than dialogue in Renaissance drama. In other words, the script always assumes that Speaker A will address Speaker B who addresses Speaker C, etc, and therefore inaccurately assigns the receiver when Speaker B responds back to Speaker A. Here, methodology illuminates dramatic structure. The linear model of "progressive" dialogue correctly attributes a higher percentage of utterances (and words) in Greek than Shakespearean tragedy. Dialogue in ancient plays more strictly observes a recursive model that occurs less frequently within Shakespeare's more populated scenes. The script correctly identifies more receiver tags in plays with smaller casts (i.e., classical drama) and scenes with two interlocutors, such as Iago and Roderigo in *Othello* I.i. When there are more than two interlocutors, the script misattributes a response to a previous speaker as an address to the following speaker. However, accuracy is usually measured at the unit of the utterance, or speech-tag. When measuring accuracy by words, Act I of *Othello* presents a different case. Word-based accuracy is consistently lower than utterance-based accuracy in the Greek plays and in *Hamlet*, but in *Othello*, the reverse is true. Thus, inaccurately attributed speeches —those given in response to a preceding speech-- are longer in *Hamlet* and the Greek plays than in the opening act of *Othello*. It follows that in *Othello*, characters tend to say more when they address the next character who speaks. If there's more chain-like dialogic progress in Act I of *Othello* than in all of *Hamlet*, this corroborates scene-based thematics that

characterize *Othello*: its beginning *in medias res* and the transmission of news through increasingly public scenes (from “a street” (I.i) to “another street” (I.ii) to “a council-chamber” (I.iii)). In the second section of this paper I explore such network patterns diachronically: at levels of scene, act and play, and in comparative context.

This formal figuration of dialogue’s progress supplements the paper’s central findings: weighted network visuals, at levels of the play entire, act and scene. At the most macroscopic level, we observe “synchronic” patterns and discontinuities: for example, larger scenes, with more characters and therefore usually a more “public” nature occur near the beginning and the ending of *Hamlet*, *Lear*, and *Othello*. Networks create a hierarchy of centrality among characters. This model inevitably calls into question the binaries with which we usually think about characters: protagonist vs. minor characters, or "round" vs. "flat": nothing here supports these dichotomies, and, in fact, the hierarchical re-conceptualization of characters is another promising research area opened by network theory.

The third section focuses not on character as determined by location in narrative structure, but by semantic analysis of dialogue, using word frequency and semantic field analysis. For example, relativized word frequency lists reveal that “Cassio” is among the top ten most frequent words that Brabantio, the Duke, Lodovico and Montano speak, before any other character’s name, including “Othello,” “Desdemona,” or “Iago.” This semantic analysis detects the linguistic virulence of Iago’s revenge scheme, “promoting” Cassio’s name in the constructed adultery plot. Semantic analysis and weighted word networks also reveal that antagonists’ confrontations are consistently verbose in Greek tragedy, while Shakespearean antagonists may never exchange a word. In this paper, I consider dramatic plot as a function of network connections--coded as relations between speaker and addressee—through both historical and generic developments.

References

Albert-László Barabási, (2003). *Linked.* NY: Penguin.

Mark Granovetter,. 'The Strength of Weak Ties'. *American Journal of Sociology.* May 1973.

R. Alberich, J. Miro-Julia, F. Rosselló. 'Marvel Universe looks almost like a real network'. `http://arXiv:cond-mat/0202174v1`.

M. E. J. Newman (2006). 'Finding community structure in networks using the eigenvectors of matrices'. *Physical Review E 74.* 036104.

M. E. J. Newman (2003). 'The Structure and Function of Complex Networks'. *SIAM Review.* .

Mark Newman, Albert-László Barabási, Duncan J. Watts (eds.) (2006). *The Structure and Dynamics of Networks.* Princeton UP.

Mark Steyvers, Joshua Tenenbaum (2005). 'The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth'. *Cognitive Science.* .

James Stiller, Daniel Nettle, Robin I.M. Dunbar (2003). 'The small world of Shakespeare's plays'. *Human Nature.* vol. 14, no. 4.

PAPER 3

Social Connections and Space in Nineteenth-Century Rio de Janeiro

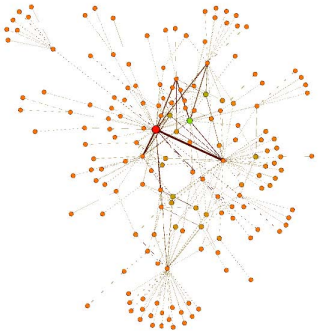
Frank, Zephyr
zfrank@stanford.edu
Stanford University

This paper explores the theme of social connections in the novels of Machado de Assis, José de Alencar, and Aluísio Azevedo against a background of empirical research on historical social networks in the space of the city of Rio de Janeiro. It argues that a better understanding of the meaning and configuration of social networks in history can be obtained through creative use of literary sources. In this, the paper builds upon the work of Antonio Candido, Roberto Schwarz, and Raymundo Faoro—all of whom emphasized the historical and sociological richness in the works of the novelists under consideration. By emphasizing the physical space of the city, its street networks and modes of transport, and the existence of overlapping social networks therein, the paper connects sociologically inflected literary criticism to social history through the use of new digital methods of analysis and visualization.

The body of the paper is divided into two parts. The first part analyzes networks within the novels and the novelistic space of the city. The second part, building

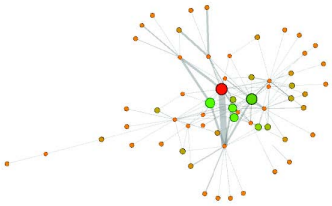
on the network typologies and spaces discovered in the analysis of the novels, explores social networks and the space of the city through the analysis of historical documents such as lists of club members or occupational groupings. The paper then concludes with an appraisal of the way literary networks and spaces can inform better historical questions put to more traditional historical documents.

In terms of digital humanities methods, the paper explores the use of computational techniques and programs such as GIS (ArcGIS and related software) and Gephi, a network analysis tool. Part of the novelty of the approach considered in this paper is the degree to which it combines analysis across these platforms —that is, network space (Gephi) and geographical space (ArcGIS) taken together. Because it is difficult to describe the approach taken in the paper in words alone, two examples of the kinds of visualization generated for the paper are shown to the right.



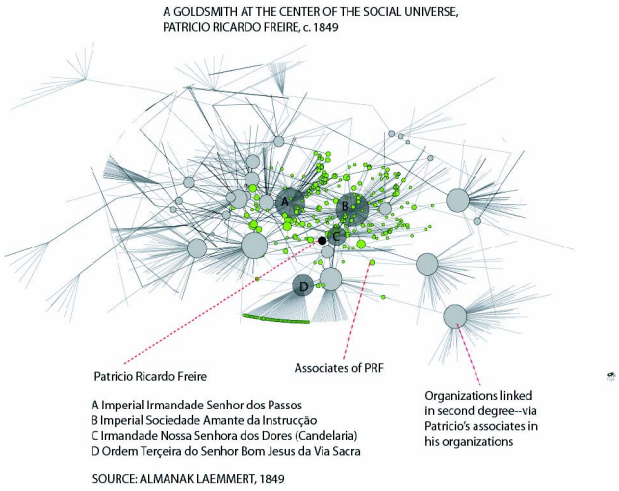
Epitaph of a Small Winner

Machado de Assis's novel (1881) is told in the first person by a voluble narrator. Connections to the central character (red node) shoot off in multiple directions as the book zig-zags through his life (which can be read, loosely, as a Bildungsroman of Brazil personified in the character of Brás Cubas). The structure of the novel thereby reflects the author's critique of the capricious and fragmented world of Rio de Janeiro. Still, in the midst of all this fragmentation, certain characters and places emerge as brokers and spaces of social connections.



Sonhos d'Ouro

José de Alencar's novel (1872) is told in a traditional third-person style. The central character (red node) is surrounded by characters of similar importance (measured by network centrality). There are relatively few network clusters and far fewer characters and complications than in the case of Machado's novel. The plot of Alencar's novel seeks to resolve the tension between newly minted capitalist wealth and values of thrift and honor embodied in the protagonist.



The second part of the paper explores what happens when we place historical individuals in the context of social networks and the spaces of the city. Using information gleaned from the analysis of networks in novels, including the role of "brokers" and "spaces of interaction," part two looks for evidence of similar patterns of social connections and spaces in the historical record.

As in the example provided in the diagram above, it is possible to reconstruct the shape of social networks in nineteenth-century Rio de Janeiro through the collection of data concerning club and organization

membership and the analysis of large datasets with tools such as Gephi and Pajek (used above). There is a clear pecking order in the organizations of the civic fauna in Rio. Just a few leaders are required, for example, to tie together 90 percent of the civic fauna in first- or second-degree connections. The same emphasis on hierarchy and gatekeeping with respect to social connections appears in the analysis of the novels discussed with respect to the first part of the paper.

The "#alt-ac" Track: Digital Humanists off the Straight and Narrow Path to Tenure

Nowviskie, Bethany
bethany@virginia.edu
University of Virginia

Flanders, Julia
julia_flanders@brown.edu
Brown University

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

Reside, Douglas
dougreside@gmail.com
New York Public Library

Porter, Dorothy (Dot)
dot.porter@gmail.com
University of Indiana

Rochester, Eric
err8n@virginia.edu
University of Virginia

1. Participants and Presentations

Dr. Bethany Nowviskie (moderator): "DH in the *#alt-ac*ademy Project: Editing *Alternate Academic Careers for Humanities Scholars*"

Director, Digital Research & Scholarship (Scholars' Lab, UVa Library); Associate Director, Mellon Scholarly Communication Institute, University of Virginia

Dr. Julia Flanders: "Accounting for Time and Labor in the Knowledge Work of the Digital Humanities"

Director, Women Writers Project; Associate Director for Textbase Development, Scholarly Technology Group, Brown University

Dr. Tanya Clement: "Off the Tracks: Laying New Lines for Digital Humanities Scholars (Results of an NEH Workshop)"

Associate Director, Digital Cultures and Creativity, Honors College; Research Associate, Maryland

Institute for Technology in the Humanities (MITH), University of Maryland, College Park

Dr. Doug Reside: "Of Ant-Lions and Scholar-Programmers: DH Centers as Ideal Habitat?"

Digital Curator for the Performing Arts, New York Public Library

Dot Porter, MLIS: "Credential-Creep in the Digital Humanities Job Market"

Associate Director for Digital Library Content & Services, Indiana University

Dr. Eric Rochester

Senior Developer, Digital Research & Scholarship (Scholars' Lab, UVa Library), University of Virginia

2. Panel Abstract

This is a panel session proposed by six established non-tenure-track practitioners and scholars of the digital humanities. Several of us are among the contributors to a forthcoming open-access essay collection on the place of "alternative academics" within the academy. Our goal is to open a discussion of issues from that collection, and from a related NEH-funded workshop on career paths in DH centers, that are relevant to the lives of digital humanists working outside the professoriate.

The "#alt-ac" project, to be published in 2011 (online, for comment and extension) by NYU's *MediaCommons*, features contributions by and for scholars with deep training and experience in the humanities, who are working or seeking employment — off the tenure track — within universities and colleges, or in allied knowledge and cultural heritage institutions such as museums, libraries, academic presses, historical societies, and governmental humanities organizations. For reasons ranging from the professional, intellectual, or institutional to the deeply personal, an increasing number of these people identify themselves as members of the digital humanities community.

The work of academic and cultural heritage institutions has long been enriched and enabled by humanities scholars, developers, and administrators like us. Regardless of institutional status, digital humanists maintain sophisticated research, publication, and production profiles and bring methodological and theoretical training to bear on problem-sets of great importance to traditional humanities disciplines and to the scholars who operate within them. However, class divisions between faculty and staff in higher education are profound, and the suspicion or (worse) condescension with which so- called 'failed academics' are sometimes met can be disheartening. Working, as we are often perceived to do, on the margins of C. P. Snow's "two cultures" (or, less polemically, at the interstices of libraries or DH centers and academic departments) digital humanities practitioners are sometimes placed well outside more easily-classified and socially-supported areas of the academic enterprise.

For all that, non-tenure-track digital humanists love their work. Many people navigating ill-defined alternative academic career paths speak compellingly about the satisfaction of making teams (and systems, and programs) work, of solving problems and personally making or enabling breakthroughs in research and scholarship in their disciplines, and of contributing to and experiencing the life of the mind in ways they did not imagine when they entered graduate school. Essays in our *#alt-ac* collection range from personal narratives, positioned within certain academic disciplines and institutions, to staged dialogues on opportunities and pitfalls off the tenure track, to reflective and data-driven essays on the state of academic digital humanities and the (problematic? disruptive? salutary?) position of "alternative academics" within it. A few contributors also illustrate retrograde career paths or offer critiques of the *#alt-ac* concept. Conversations stemming from the MediaCommons project have sparked a number of other collaborations, such as MITH's "Off the Tracks," an NEH workshop on institutional practices in "laying new lines for digital humanities scholars" — preliminary recommendations from which will be presented here.

The six speakers on our proposed panel will offer 7-minute summaries of their research or make position statements based on their published essays, before opening discussion amongst themselves and with the DH conference audience.

Briefly, the panelists include:

Dr. Bethany Nowviskie of the University of Virginia, who will also moderate the discussion. Nowviskie, who speaks and writes frequently on alternative academic careers in the digital humanities, is the editor of the forthcoming MediaCommons collection. She will frame this panel's discussion with an overview of concepts relevant to the digital humanities that have emerged in all contributed essays and dialogues and will reflect on some issues crucial to DH — such as intellectual freedom and policies related to open source and intellectual property — that were note foregrounded as clearly in the collection as one might expect.

Dr. Julia Flanders of Brown University. Flanders will speak to issues raised by her dissertation on the politics of labor in the digital humanities and her #alt-ac essay, "You work at Brown — What do you Teach?," which examines the pragmatic and psychological effects of divergent practices in accounting for time and labor in knowledge work. How do we think of — and reward — time and effort differently when laborers on very similar projects may alternately be classified as adjunct, tenured, or tenure-track faculty, salaried staff, hourly wage employees, student apprentices and fellowship winners, and post-docs or research staff? What are the implications for our institutions and our digital humanities workforce?

Dr. Tanya Clement of the University of Maryland, College Park. Clement will describe the genesis, process, and outcomes of a two-day NEH workshop held at MITH, the Maryland Institute for Technology in the Humanities, in January of 2011. "Off the Tracks: Laying New Lines for Digital Humanties Scholars" will examine best practices across a number of differently-constituted digital humanities labs and centers, with the goal of articulating reasonable career trajectories and professional development opportunities for hybrid academics and scholar-programmers — employees who are not well served by a normative division between "research" usually associated with faculty positions and "service" usually associated with staff roles.

Dr. Doug Reside of the New York Public Library. Reside, formerly a "scholar-programmer" at MITH, is Clement's partner in the NEH workshop. His presentation will be informed by that experience and their collaborative drafting of a white paper based on the workshop, but he will also speak to the themes of his #alt-ac essay, "Of Ant-Lions and Scholar-Programmers." Here, he argues that — of several imperfect options for satisfying the two (unified but sometimes contradictory) natures of the digital humanities software developer, a well-defined position in a digital humanities center holds the most promise.

Dot Porter, MLIS of the University of Indiana. As a distinguished member of the DH community and the only panelist not holding a doctorate in the humanities, Porter will speak to the issue of "credential creep" in the digital humanities job market. Her #alt-ac essay, co-authored with Amanda Gailey of the University of Nebraska, likens the frequency with which DH jobs are now advertised as requiring a doctorate to the conditions described in William James's 1903 essay, *The PhD Octopus*. Gailey and Porter examine increasingly stringent job pre-requisites and what they term the "creeping Ivy" of elitism in the context of traditional spirits of entrepreneurship and egalitarianism in the digital humanities. Porter will conclude by offering clear recommendations to DH hiring committees.

Dr. Eric Rochester of the University of Virginia. After getting a PhD in English from the University of Georgia, Rochester worked as a computational linguist and programmer for a number of technology firms. He joined the Scholars' Lab at the University of Virginia Library as senior software developer in 2011. His presentation, “There and Back Again,” will discuss transitioning between industry and academia, and how to manage and ease that process for both the individual and the institution.

All six speakers have enthusiastically committed to attend and present at DH in Stanford. We recognize that the presence of "alternative" academic career paths is long-established in the digital humanities — if in generally ad-hoc ways — and will not argue that our roles and observations are fundamentally new. However, we feel strongly that this conference — with its focus on the "big tent" of the digital humanities, and occurring at a time when humanities graduate students are faced with the worst tenure-track job market in memory — is the right moment for a sustained and critical discussion of the opportunities and challenges of DH career-seekers off the straight and narrow path to tenure.

References

James, W. (1903). 'The Ph.D. Octopus'. *Harvard Monthly.* 36: 1-9.

Nowviskie, Bethany (ed.) (2011). *#alt-academy: Alternative Academic Careers for Humanities Scholars.* MediaCommons: a digital scholarly network. `http://mediacommons.futureofthebook.org/`.

Off the Tracks: Laying New Lines for Digital Humanities Scholars. `http://mith.umd.edu/offthetracks/` (accessed 19-20 January 2011).

The Social Networks and Archival Context Project

Pitti, Daniel
dpitti@Virginia.edu
Institute for Advanced Technology in the Humanities, University of Virginia

Larson, Ray
ray@ischool.berkeley.edu
University of California, Berkeley. School of Information

Janakiraman, Krishna
krishna.j@berkeley.edu
University of California, Berkeley. School of Information

Tingle, Brian
brian.tingle@cdlib.org
University of California, Berkeley. School of Information

This session will present an interim report on the findings and a demonstration of the *Social Networks and Archival Context* (SNAC) project.

SNAC is exploring the feasibility of using existing archival descriptions to create a prototype socio-historical resource and resource discovery tool that will enhance access to and understanding of cultural resources in archives, libraries, and museums. Beginning in May 2010, with funding from the National Endowment for the Humanities, the two-year project is using advanced technology in three primary ways: to derive descriptions of people from descriptions of their records; to match and merge the derived records with library and museum authority records; and to build a prototype socio-historical resource and access system use the resulting matched and combined records.

Leveraging the new standard Encoded Archival Context–Corporate Bodies, Persons, and Families (EAC-CPF), the project is deriving EAC-CPF records from nearly 30,000 EAD-encoded finding aids made available to the project by the Library of Congress and three archival consortia: Online Archive of California; Northwest Digital Archive; and Virginia Heritage. Names of record creators and of people documented in the records and record descriptions are being used to create EAC-CPF records. Co-occurrence of names in finding aids is also being recorded in order to document social and professional relations among named entities and to interrelate (or link) the related records with one another. The resulting EAC-CPF records are matched against one another and against several million authority records represented in the Library of Congress Name Authority File (NACO/LCNAF), the Getty Vocabulary Program's Union List of Artist Names (ULAN), and the Virtual International Authority File (VIAF), a collaboration between several national libraries. Unique data in matching records is being merged or combined into a single EAC-CPF record, with the ultimate goal of having one record for each unique person, corporate body, for family. Finally, the project is developing a prototype public access and historical resource system based on the unique EAC-CPF records created from the processing.

SNAC is a collaboration between three institutions: the Institute for Advanced Technology in the Humanities (IATH), University of Virginia; the California Digital Library (CDL), University of California; and the School of Information, University of California, Berkeley (SI/UCB). IATH is the lead institution and is responsible for overall project management and for deriving EAC-CPF records from EAD-encoded finding aids. SI/UCB is responsible for the matching and merging of authority records. CDL is responsible for developing a prototype public access and historical resource system based on the data produced by the other two partners.

The three papers presented in the session will cover the following topics: A comprehensive overview of SNAC and a detailed description of the derivation processing; a description of the theoretical and application challenges of matching data from heterogeneous sources; and a description of the methods and technology being adapted or developed to create the prototype public system. Finally, the session will present a brief demonstration of the prototype system.

PAPER 1

Overview and Methods

Pitti, Daniel
dpitti@Virginia.edu
Institute for Advanced Technology in the Humanities, University of Virginia

1. Introduction

Archivists have a long history of describing the people who—acting individually, in families, or in formally organized groups—create and are documented in archival records. They research and describe the artists, scholars, political leaders, scientists, government agencies, soldiers, universities, businesses, families, and others who create and are represented in the items that are now part of our shared cultural legacy. However, because archivists have traditionally described records and the people documented in them in a single apparatus, the finding aid, this biographical-historical information is tied to specific resources and institutions. Currently there is no system in place that aggregates these descriptions of individual persons, families, and corporate bodies, and interrelates them to reveal the professional and social relations that existed among the described entities.

Leveraging the new standard Encoded Archival Context–Corporate Bodies, Persons, and Families (EAC-CPF), the *Social Networks and Archival Context* (SNAC) project is exploring the feasibility of using existing archival descriptions in new ways in order to enhance access to and understanding of cultural resources in archives, libraries, and museums. Beginning in May 2010, with funding from the National Endowment for the Humanities, the two-year project is using advanced technology to derive descriptions of people from descriptions of their records, to match and merge the derived records with library and museum authority records, and use the resulting records to build a prototype socio-historical resource and access system.

2. Data and Data Contributing Institutions

Over 28,000 EAD-encoded finding aids have been made available to the project by the Library of Congress (918+) and three archival consortia: the Online Archive of California (13,932+); Northwest Digital Archive (5,160+), and Virginia Heritage (8,390+). The three consortia represent findings from over 200 individual repositories.

A key criterion in selecting the finding aid sources was geographical proximity. Because archives commonly emphasize local history, it is surmised that the geographic proximity of the archives would yield a higher rate of co-referencing (for example, a correspondent in one finding aid is the creator in another), and thus provide corroborating evidence of social-professional relations. Another very important consideration was encoding consistency, and thus the project emphasized consortia with high quality control standards.

Three institutions are contributing authority records that will be matched and merged with the derived EAC-CPF records. The Library of Congress has made the NACO/Library of Congress Name Authority File (LCNAF) available for project use (3.8M personal and 900K corporate name records). The Getty Vocabulary Program has contributed the Union List of Artist Names (ULAN) (293K personal and corporate names). Finally, OCLC Research has made available the subset of the Virtual International Authority File (VIAF) that intersects with the LCNAF available to the project. The VIAF, for now, only contains personal names.

3. Separating Description of People from Description of Records

There are several interrelated intellectual and practical rationales for separating the descriptions of people from the description of the records that document their lives. These rationales are based on archival processing efficiencies, the intellectual quality and depth of resource description, and enhanced access to primary humanities resources for all users.

Cooperative Authority Control. Authority control is labor-intensive, and sharing the creation, maintenance, and use of authority data improves catalogers' productivity. Sharing descriptions of creators and people and organizations documented in archival records saves time and labor.

Integrated Access to and Context for Cultural Heritage. Integrated, union access to archival authority records can be used to locate and identify people, organizations, and families, and these records in turn can lead to cultural heritage resources through links to descriptions or dynamically generated searches of catalogs. An archival authority record can provide not just contextual information for understanding archival resources, but also access to and context for

understanding all that constitutes the human record and our cultural heritage.

Biographical/Historical Resource. In addition to name control, archival authority records provide biographical/historical data about the creator, such as when and where the creator existed, significant activities and functions performed by the creator, and other significant dates, places, and events. This historical information can be used as an independent resource that can assist users in identifying and learning about the described entity.

Social/Historical Context. People live and work with other people, both as individuals and as members of families and organizations. These social and professional relations are reflected in records created by them and consequently in the descriptions of the records. Archival authority control records provide a potential means to systematically gather and document these social and professional relations in links that interrelate descriptions of people, organizations, and families. This documentation can provide convenient access to the broad social-historical contexts within which corporate bodies, persons, and families were active, and convenient, navigable access to related or complementary resources.

4. Methods and Processing Overview

There are three ways in which technology is being used in the SNAC Project. First, using Extensible Markup Language–Transformation (XSLT), the project is deriving EAC-CPF records from EAD-encoded finding aids for record creators and people referenced in the description [1]. Initially the project is focusing on extracting names, biographical-historical data, occupations, dates of existence, and languages used that are clearly and specifically encoded in the finding aids. Later in the project, natural language processing techniques will be used to experiment with extracting names and other targeted descriptive data that appear in the description but that are not encoded specifically. Second, the extracted EAC-CPF records are being matched against one another and against NACO/Library of Congress Name Authority File (LCNAF) records, Union List of Artist Names (ULAN) authority records, and finally Virtual International Authority File (VIAF) aggregated authority records. Unique data in matching records is being merged or combined into a single EAC-CPF record. Finally, the project is developing a prototype public access and socio-historical resource system based on the collection of unique EAC-CPF records created and interrelated.

This paper will address the first step in the processing, deriving EAC-CPF records from archival finding aids. The matching and merging and development of the prototype system will be addressed in separate papers.

5. Deriving EAC-CPF Records

The principal technologies involved in the derivation process are XSLT 2.0 and XPath 2.0, with relatively heavy use of regular expressions and customized functions. Initially in the project, the focus is on identifying and deriving individual records from the following EAD tag components: <persname>, <corpname>, and <famname> that occur within <origination>, <controlaccess>, and <unittitle>. Personal, corporate, and family names derived from <origination> and <controlaccess> are generally formulated according to strict cataloging rules (AACR2, for American archives and libraries), though challenges are presented by names that are poorly formulated (for example, in direct rather than inverted order, and intermixed with non-name data, subject subdivisions or uniform titles). While many names occurring within <unittitle> are tagged as such, many occur without being tagged as names. The tagged names found in <unittitle>, like those found in <controlaccess>, are irregularly formatted. Regular customized functions and named templates, many incorporating the use of regular expressions, are used to isolate and normalize the name strings, and to create unique lists of names.

For each unique name string found, an EAC-CPF record is created. For records derived for creators, additional descriptive data for dates of existence, occupation, subject headings assigned to records, languages used, and biographical-historical information is extracted into the corresponding EAC-CPF records. Additionally, all unique referenced names are related to the EAC-CPF record for the creator, and for each an EAC-CPF record is also created, and related to the record for the creator. Because of co-referencing among finding aids (for example, the same person corresponded with two more record creators), the resulting set of records derived from any set of finding aids contains more than one EAC-CPF record describing the same entity. Thus while duplicate entries are not created for named entities found in a particular finding aid, duplicate or, more accurately, matching records are created in the processing of all of the finding aids.

6. Conclusion

While SNAC has only been underway for five months, the Library of Congress, Online Archive of California, and Northwest Digital Archive finding aids have been processed to derive nearly 160,000 EAC-CPF records. The matching and merging processing at SI/UCB began in September 2010, after several weeks devoted to acquiring and indexing several million VIAF, ULAN, and LCNAF authority records in preparation for the matching and merging processing. The initial release of the prototype public historical resource and access system was in December 2011. Though many challenges remain, the early results suggest that the data and the methods and techniques being applied are highly effective. The deriving, and matching and merging processing will continue to be refined, and the prototype public system will continue to be refined and new features developed.

Notes

1. The co-occurrence of names in the description of a single collection documents either a social-professional or intellectual relation between the named entities. Some occurrences can specifically be identified as correspondents, thus confirming a social-professional relation.

PAPER 2

Matching and Merging EAC-CPF Records

Larson, Ray
ray@ischool.berkeley.edu
University of California, Berkeley. School of Information

Janakiraman, Krishna
krishna.j@berkeley.edu
University of California, Berkeley. School of Information

1. Introduction

Our interests in cultural heritage, history, and the social sciences are fundamentally about human activities. Understanding the circumstances of people's actions —who, what, when, how, and why—illuminates their lives and the events that they experienced. While much information of interest to scholars is already available in the collections of cultural institutions such as archives and libraries, there is a significant gap in the information infrastructure for dealing with information about people. Standards for the computerized handling people's names have been developed in libraries (such as MARC Authorities). With the development of the EAC-CPF (Encoded Archival Context–Corporate Bodies, Persons, and Families) standard, a similar capability is just now becoming available to archival collections. The Social Networks and Archival Collections (SNAC) project aims to start bridging that gap and to connect the information about corporate bodies, persons and families in the library world with those entities in the archival world.

This paper reports on current experiments in matching and merging entities in collections of EAC-CPF records with those in library authority files and other sources. EAC-CPF records represent the entities (which can be individuals or groups of individuals) mentioned in archival description records and can be derived from the EAD (Encoded Archival Description) records that encode description of archival collections. EAD records, created by archivists and librarians, serve as vital finding aids. Information on entities in these records is an invaluable reference for humanities scholars, particularly since entities may be referenced and represented in multiple archival descriptions.

EAC-CPF records encode extensive information about an entity, drawn from various parts of the source records. In addition to basic identifying information (name, type, occupation(s), and existence dates), they include an entity's relationship(s) with other entities, resources, and works. Since EAC-CPF records are derived independently from each EAD record, there can be multiple records representing the same entity in multiple EAD collections. A key problem, then, is to identify multiple EAC-CPF records that represent the same entity and merge them together into a single record.

SNAC has been given an extensive collection of library Name Authority records, including the Library of Congress Name Authority File (LCNAF), the Virtual International Authority File (VIAF) from OCLC Research, and the Union List of Artists Names (ULAN) from the Getty Vocabulary Program. The VIAF database combines name authority files from a number of libraries worldwide, including the Library of Congress, La Bibliothèque nationale de France, Deutsche Nationalbibliothek, and the Vatican Library. SNAC's current implementations use either an exact string match criteria or the alternate name information from the name authority files to match entities in the EAC-CPF collection.

2. Related Work

Our problem is similar to the well-studied entity name disambiguation problem, where the task is to identify the correct entity, under a given context, from a set of seemingly identical entities. Standard approaches use statistical learning techniques, either performing supervised learning and train classifiers that predict the relevance of an entity given a context or performing unsupervised learning and design clustering techniques that cluster similar entities together. As an example of the former, Bunescu and Pasca (Bunescu and Pasca, 2006) suggest a method that trains Support Vector Machines (SVM) classifiers to disambiguate entities using the Wikipedia corpus. The classifier was trained using features extracted from the title, hyperlinks linking other entities, categories assigned to the entity and Wikipedia's redirect and disambiguate pages. Bagga and Baldwin (Bagga and Baldwin, 1998) and Mann and Yarowsky (Mann and Yarowsky, 2003) are examples of the latter technique, where similar entities are clustered using features extracted from entity's biographical information, words from sentences surrounding the entity in texts and entity's social network and relationships. Other techniques involve using gazetteers and name authority files as external references to aid the disambiguation process. Smith and Crane (Smith and Crane, 2001), for example, use gazetteers to disambiguate geographic place names.

SNAC's focus is more precisely a clustering problem rather than a classification problem, since we want to group, in an unsupervised way, EAC-CPF records belonging to the same entity. While some of the work mentioned above uses sophisticated techniques to discover entities in the text, the EAC-CPF standard provides direct access to the name and other information about the entity. This, combined with the availability of the name authority files, allows use of simpler algorithms based on exact string match and authority file look up for matching entities.

3. Implementation

We have 158,079 EAC-CPF records — 114,639 persons, 41,177 corporate bodies and 2,263 family names — derived from Library of Congress, the Online Archive of California, and Northwest Digital Archive EAD records. The records were parsed with the EAC-CPF specification to extract information on an entity's name, type, and relations, stored in a relational database. Preferred and alternate names from the VIAF name authority files were indexed using the Cheshire II information retrieval engine (Larson et.al, 1996), which uses a probabilistic information retrieval algorithm to find the top n VIAF records and their associated names given an entity name. We mapped each EAC-CPF entity to names from top n VIAF records (currently the top five VIAF records). These mappings are also represented in a relational database.

The primary approach is based on the simple hypothesis that two EAC-CPF records belong to the same entity if the entity names exactly match. This simple technique reduced the total number of unique records to 129,915. This meant that EAC-CPF records with different names belonging to the same entity were not matched. This was often due to the presence of existence dates in the name fields (e.g., "Einstein, Albert, 1879-1955" will not match "Einstein, Albert").

Our second technique uses the hypothesis that entities referring to the same name authority record must be the same. Our database and index setup made this easy. For a given pair of entities, we search the entity VIAF names mapping table for alternate names for the first entity. If the second entity's name appears in the first entity's list of possible alternate names, we consider the two entities to be the same. Match results are parameterized by how liberal the search is: including names from all the top five ranked VIAF records would result in a higher number of matches but fewer accurate matches. Using names from higher ranked VIAF records would give a lower number of matches with better accuracy. However, evaluation of the technique showed that using the best matching or top ranked VIAF record reduced the number of unique EAC-CPF records to 124,657, or 5248 records less than what was achieved using the exact name match technique. This was contrary to our expectations and further evaluation suggested that it was a result of either subtle differences in the way the names are spelled and punctuated or the use of names that are not present in the authority files. Using lower ranked VIAF records reduced the number of unique records but introduced serious matching errors and was not a viable option.

4. Conclusions

We have described our current implementations that match and merge EAC-CPF records belonging to the same entity. Our implementations use exact name match as the criteria or use the name authority files as an external reference to disambiguate and find matches. Our current technique finds mostly accurate matches; false matches occur only when there are two

different entities with the same name or when the name authority file has inaccurate information. Although it is possible for different entities to have the same name, the use of existence dates can differentiate the names and given that the VIAF record combines information from institutions with rigorous standards, it is unlikely that the records will have inaccurate information.

However, our current approach still fails to identify many possible matches. The main reasons seem to be subtle variations in names and punctuation or use of names that are not present in the authority files. To handle spelling issues, we plan to experiment on using string comparison algorithms (such as “edit distance” algorithms) and use comparison results as features for a clustering algorithm. We will also experiment with other information about the entities, such as their biographies, relations with other entities, works produced, etc., and external sources such as DBPedia. However, because entity description records are created for individual archives, it is possible that this additional information is non-redundant and therefore not useful for matching purposes.

References

Bagga, A., Baldwin, B. (1998). 'Entity-based Cross-document coreferencing using the vector space model'. *Proceedings of the 17th International Conference on Computational Linguistics.* V. 1, pp. 79-85.

Bunescu, R., Pasca, M. (2006). 'Using encyclopedic knowledge for named entity disambiguation'. *Proceedings of EACL.* V. 6.

Larson, R. R., McDonough, J., O’Leary, P., Kuntz, L., Moon, R. (1996). 'Cheshire II: Designing a next-generation online catalog'. *Journal of the American Society for Information Science.* 7: 555-567.

Mann, G. S., Yarowsky, D. (2003). 'Unsupervised personal name disambiguation'. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003.* V. 4, pp. 33-40.

Smith, D., Crane, G. (2001). 'Disambiguating geographic names in a historical digital library'. *Research and Advanced Technology for Digital Libraries.* Pp. 127-136.

PAPER 3

Paper Three: The Social Networks and Archival Context Project: Developing a Prototype Historical Resource and Access System

Brian Tingle
brian.tingle@cdlib.org
University of California, Berkeley. School of Information

1. Introduction

While authority file interfaces for library authority control is reasonably well understood, archival authority records provide more detailed description and possibly extensive entries to related persons, corporate bodies, and families, and related resources. This paper will explore the opportunities and challenges faced in designing and implementing a public interface to access the unique archival authority record aggregations created by the SNAC project.

The prototype access system [1] strives to enable humanities researchers to make use of archival context records that describe individuals, families, and formally organized groups to find resources and identify social and professional relations that are hard to discover using existing research tools and techniques. Existing methods require searching and exploring dispersed archival finding aids and using the descriptions found to further search and explore dispersed finding aids. The SNAC access system does this work for the researcher, bringing this information together in a searchable database and exposing the discovered social graph as open linked data.

The prototype was made available for public testing on December 17, 2010. From December 17, 2010, through February 17, 2011, the site had 47,857 visits and 104,223 page views, with 91.37% of traffic coming from Google searches. An iterative development model is being utilized, and user feedback on the prototype will help us to identify and prioritize ongoing development activities. New features are released into the prototype site on an ongoing basis, as they are developed. Grant-funded interface development will cease in April 2012 with the completion of the project.

Source code for the access interface is also available as open source software [2].

A copy of the social graph derived from the research data was released as a graphML file [3] on February 17, 2011, under the Open Data Commons Attribution License [4]. Analysis is ongoing to figure out the optimal way to represent the social graph in a conventional RDF model (such as with the FOAF ontology).

2. Technical Infrastructure

The semantics and structure of the records based on the EAC-CPF schema has been the starting point of the initial prototype. The prototype is being built as an application of the eXtensible Text Framework (XTF) and will support search, display, and navigation of EAC-CPF records [5]. The Tinkerpop Graph Processing Stack [6] is also being utilized by the project to load the relationships recorded in the derived EAC records into a graph database and exposing it though a REST API to the interactive graph visualizations in the prototype interface. The Tinkerpop stack is compatible with linked data technologies through the RDF Sail interface, and will be used as the platform to expose the project's data as linked data when the semantic modeling is complete.

3. Current and Future Development

The access system being developed resembles library systems based on authority control, but the EAC-CPF archival authority records are far more complex than those created by the library community. In addition to entry control (authorized and alternative names), the archival authority records frequently have biographical-historical data, occupation, dates of existence, languages used, as well links to related people and resources. An additional challenge is presented by relative quantities of descriptive data found in each record. Some records have as many as 50 or more alternative names, scores of subject headings, more than 50 related persons, families, or corporate bodies, and many linked archival finding aids or titles. Other descriptions are quite brief, based on the name occurring in one finding aid and failing to match an authority record. Finding the right method for displaying and facilitating navigation of this data presents many challenges.

The initial prototype focuses on searching, browsing and displaying the EAC-CPF records as formatted web pages for researchers. Both full text (description less control data) and specific component searches are supported. Full text searches are weighted to give preference to matches in the <identity> section of the record, where all forms of the name of the entity discovered in the derivation, matching, and merging processes are listed. Limiting a search to the <identity> section restricts the retrieval to just the forms of name found in the section, and thus excludes matches in other parts of the description, such as in entries for related named entities. As users enter searches, authorized forms of names are suggested. Users can browse the top occupations and subjects in addition to an alphabetical index of all names in the database. The alphabetical index feature is likely to become less useful as the number of records increases. (At this date, there are over 123,920 named entities.) Search results can be narrowed down to names that have a particular occupation or subject term associated, or restricted to entity type, such as person, corporate body, or family.

In the initial prototype, a wide variety of data and links are displayed to the user. For each EAC-CPF record, the following descriptive components are displayed: authoritative name, alternative names, dates of existence, sex, affiliated countries, occupations, subject terms used in describing related archival records (for record creators), and biographical/ historical description (either as prose or a chronological list). In addition to the above biographical/historical data, the following linked information will also be displayed: related persons, corporate bodies, and families; descriptions of related archival records (that is, finding aids within which the name was discovered), published work by or about the described entity. Links are also provided to a matching Virtual International Authority File (VIAF) record, when one is identified as matching.

Though the outbound links to finding aids and VIAF records are currently implemented, the internal "links," for now, are implemented as searches.

While the derivation, matching, and merging processing continues, the persistence of any given EAC-CPF record cannot be ensured: a record may be merged into another record in subsequent processing, so it is difficult to assign persistent identifiers or addresses in links to related entities. Once the deriving, matching, and merging is complete (late in the project), links to associated persons, corporate bodies, and families will directly retrieve the related records.

The list of titles for resources by and about the described entity that have been gathered in the record is currently inactive. We anticipate eventually using entries in this list to query WorldCat. When the project is complete, additional links will be

made to descriptions of named entities in DBpedia and WorldCat Identities, where matching entries are found. Also under consideration is offering users the opportunity to use authoritative and alternative name entries to search a selection of archive, library, and museum access systems, and public resources such as Google, Bing, and Flickr.

Another objective of SNAC is to employ a display and navigation tool to graphically display and facilitate the navigation of the social and professional networks discovered and documented in the EAC-CPF records and their relations to one another. Visualization of abstract networks is a well-studied problem and there are many tools available that support the graphML file format that the project has used to represent the historical social graph released under an open license. The project's objective is to develop a visualization and navigation interface that will make it possible for humanities researchers to explore and discover social-professional relations and related resources that would be difficult to explore and discover using simple lists. Experimentation with appropriate graph visualization and navigation interfaces is ongoing at the time of this writing, and a primitive version of this feature was released in February 2010.

In the public interface to the prototype resource and access system, we are concerned not only with people interacting with the system; but also their computational agents. The semantic web and open data movements are developing new ways of linking and expressing the relations between data used by researchers. The project intends to make the EAC-CPF records available as Linked Data. At a minimum, the XML files will be made available, but the project will also explore using RDF and perhaps other mappings of all or some of the data.

At the time of the conference the prototype resource will still be under active development. The paper will describe and demonstrate the latest version of the prototype available at the time of the conference, and solicit feedback from the participants that will be helpful in setting the final development agenda for the final year of the project.

Notes

1. http://socialarchive.iath.virginia.edu/xtf/search
2. http://https://bitbucket.org/btingle/cpf2html
3. http://https://code.google.com/p/eac-graph-load/downloads/detail?name=eac-graph-load-data-2011-02.tar
4. http://www.opendatacommons.org/licenses/by/
5. http://xtf.cdlib.org/
6. http://markorodriguez.com/2011/03/03/tinkerpop-as-of-spring-2011/

Literary Practice and the Digital Humanities, Redux: Data as/and Poetry

Raley, Rita
raley@english.ucsb.edu
Associate Professor of English, University of California, Santa Barbara

Baldwin, Sandy
clc@mail.wvu.edu
Associate Professor of English and Director of the Center for Literary Computing, West Virginia University

Montfort, Nick
nickm@nickm.com
Associate Professor of Digital Media, MIT

Wardrip-Fruin, Noah
nwf@soe.ucsc.edu
Associate Professor of Computer Science, University of California, Santa Cruz

Cayley, John
cayley@shadoof.net
Visiting Professor of Literary Arts, Brown University

This panel brings together four of the central theorist-practitioners in the field of what is commonly called electronic literature for a discussion of the question of data as/and poetry. Drawing on a range of aesthetic and discursive traditions – from sublimity to communication theory – panelists will consider the structural and operational logics of writing in programmable and networked media. As a foundation for more general theoretical inquiry, they will engage specific compositional practices ranging from Perl poetry generators to writing with and against Google search algorithms. In broad terms, then, the panel will contribute to the conversation about the role and function of literary aesthetics and practices in the digital humanities. It is in sympathy with the "material turn" in the digital humanities toward the platform and the mechanism, but it seeks more fully to understand the relations between material form and aesthetic effects (and affects).

Sandy Baldwin considers the problem of data and poetry in terms of what might be called the unaccountable or "wayward." He argues that such waywardness is historically and formally tied to poetry, situating data in the discourse of the sublime and in relation to a tradition of the metaphoric and the figural. As poetic discourse, the wayward or unaccountable is a problematic enunciative tactic, the rhetorical "lighting up" of detritus into significance and readability. Various "codework" practices can be seen as pursuing such tactics as uttering data and enunciating the author's name through the data midden. Nick Montfort discusses some of the textures of data and code, and the complexity of these two categories, in Montfort's *ppg256* (Perl Poetry Generator in 256 Characters) series. Even in the acrostic method of Jackson Mac Low, the set of "data" consists of two different sorts of data, source text and seed text, that are used differently. In the *ppg256* series, the only textual "data" is in quoted strings embedded in the code. Despite these complexities, making the distinction between code and data is important to understanding how *ppg256* generators work and is important to the poetics underlying these programs. Noah Wardrip-Fruin's presentation traces one particular birth narrative of an operational logic, that of the n-gram as a literary logic. It connects mathematicians such as Claude Shannon and Andrei Markov, the contributors to popular personal computer publications such as *Byte Magazine* and *Scientific American*, as well as figures such as Hugh Kenner and Charles O. Hartman. Finally, in contrast, the independent development of literary n-gram techniques by artists such as John Cayley is used to ask whether such logics are as different from collision detection (and others representing the physical world) as they may seem. John Cayley's remarks will be directed to those tools, instruments and services that now give us close-to-no-cost access to indexed, mapped, statistically modeled, data-driven views of the largest corpus of language practice on the planet. He will do this with reference to three ongoing projects: 'writing to be found,' which explores techniques for generating aesthetic linguistic forms with=against services like Google; *The Readers Project* (with Daniel C. Howe), which might be thought of as 'writing through visualizing reading'; and The Natural Language Liberation Front, which engages, agonistically, with the institutions and institutional consequences of all these new relations of literary production.

The Interface of the Collection

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

Ruecker, Stan
sruecker@ualberta.ca
University of Alberta

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta

Sondheim, Daniel
sondheim@ualberta.ca
University of Alberta

Radzikowska, Milena
mradzikowska@gmail.com
Mount Royal University

Organisciak, Peter
organis2@illinois.edu
University of Illinois at Urbana-Champaign

Brown, Susan
sbrown@uoguelph.ca
University of Guelph

PAPER 1

Introduction

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

Ruecker, Stan
sruecker@ualberta.ca
University of Alberta

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta

"just as interface cannot – finally – be decoupled from functionality, neither can aesthetics be decoupled from interface." (Kirschenbaum 2004, IV.34)

What is the interface to a collection? How has the interface to scholarly collections or corpora changed from print to the web? What interfaces are possible? As the scale of information that we have access to grows exponentially we are increasingly dealing with collections of documents rather than documents as individuals. These collections, whether they are craft collections of TEI documents like the Globalization Compendium (`http://globalautonomy.ca`) or industrial collections like Google Books (`http://books.google.com/`), are the way we see through to human artifacts and the way we manage them. Collections and the ways they are put together impose interfaces on the individual artifacts in order to collect them. For this reason we are organizing a panel that pays attention to the collected interface and its evolution. While the interface to the collection would seem unimportant compared to the interface for reading the artifact itself, in this epoch of digital excess we believe that it is the corpus interface that is the way into the excess and it is the corpus interface that structures the readers perceptions of the scope and purposes of any collection.

The apparently obvious distinction between data and interface is troubled in various ways, most often from the user perspective, where details of the underlying architecture are not visible, so that the interface becomes, for all practical purposes, the data. From the perspective of the developer, on the other hand, the issue is one of layers of interface, not necessarily in the graphical user sense, but rather between different forms of representation of the data and code. For certain kinds of information graphics, theorists like Ben Schneiderman point out that one particularly effective design strategy involves "direct manipulation", where the data is treated by the user as a kind of interactive widget (Shneiderman, Williamson & Ahlberg 1992). In this case, as in our final presentation in this panel, the question becomes one of how many layers of interface/data can be usefully superimposed to create not just a communicative environment, but also one where arguments can be formulated and considered.

In this panel, members of the Interface Design team of the Implementing New Knowledge Environments (INKE) project will discuss the manner in which collection interfaces are influenced by the structure of the materials included, by the history of traditions for representing collections, and by the intended use/needs of the users. We will review both print and digital interfaces in an effort to understand the way in which interacting with textual and non-textual content has changed and evolved.

As Clay Shirky points out, the problem of information overload makes us feel good as it explains why we aren't getting anything done. It is also a standard

starter for explaining the need for new information technologies. In Plato's *Phaedrus* Socrates tells the story of the invention of writing (274c – 275b) which is developed ostensibly to help our memories and therefore make us wiser. Vannevar Bush in "As We May Think" likewise uses the problem of overload to introduce his ideas about new technologies like the hypertext workstation the Memex (1945, p.102). Clay Shirky, after acknowledging how overused this trope of overload is by technology writers, goes on to use it again to argue that the problem is actually "filter failure" (2008). In this panel we too are reusing this trope of overload and overabundance to practice what Matthew Kirschenbaum calls for in "'So the Colors Cover the Wires': Interface, Aesthetics, and Usability." Kirschenbaum argues that computers are "*venues* for *representation*" (emphasis in the original) and therefore we need to study the aesthetics of their interfaces, though the aesthetics can't be detached from functionality (2004, IV.34)

This panel will, in sum, start with a general introduction of the problem, then look at the collected interface through five short interventions that either reflect back on the evolution of interface or look forward showing experimental interfaces as to what could be.

PAPER 2

The Citation from Print to the Web

Sondheim, Daniel
sondheim@ualberta.ca
University of Alberta

What holds the collection together? One of the most important interface features for scholars is the lowly citation. The citation, whether in print or hypertextually instantiated on the web, is one of the interface features that weaves together citations through associations. In this panel presentation, we will survey the design of the citation as it has evolved from print to the web by presenting a draft topology of five citation design patterns:

Absence of citation is a (non)pattern of missing citation that reminds us of the historical construction of citation and its importance to scholarship. We will compare instances of missing references in print to their online equivalents that suggest a collaborative future for citation. Perhaps all we need to connect to a reference is a search string for Google.

Juxtaposition is a pattern of citation design where the citation is placed in the flow of the text, as in inline citation. The connective value of the citation comes from what it is juxtaposed with on the page. We will expand on inline citation to look at other spatial arrangements of information, both on the page and the screen.

The **Canonical citation** is a pattern that points to an ontologically general idea of a work. We will show how canonical citation has been instantiated online such that the user can interact with multiple editions of the referenced.

The **Footnote** is a pattern of citation design that creates a relationship between source text and cited text in a visual space. The citation is moved to the foot, but it can be moved aside in other ways especially online.

The citation of **Other Media** is a pattern in which the object of citation is of a different type than the citation itself. We are seeing citations within digital video and computer games. How are we to understand the adaptation of citation expectations to other media?

To conclude, we will show that current interfaces for navigating collections bear similarities to particular forms of citation, but that navigation is beginning to evolve beyond its print patterns.

PAPER 3

The Paper Drill

Ruecker, Stan
sruecker@ualberta.ca
University of Alberta

In this panel presentation, we consider the question of how the principles of rich- prospect browsing can be used to extend the design of a database-reporting tool for journals in the humanities. While the previous presentation surveys existing interfaces, this one presents a new experimental interface, The Paper Drill, designed to navigate collections of articles through citations.

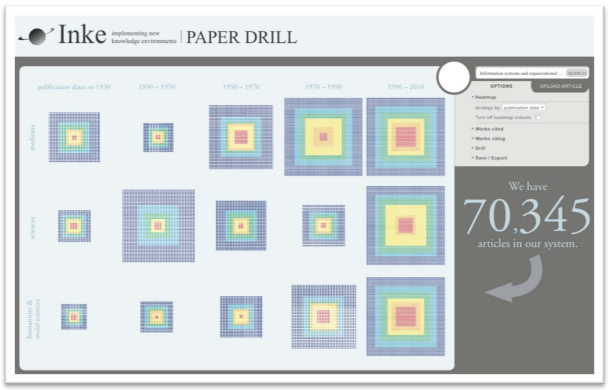


Figure 1: Paper Drill prototype

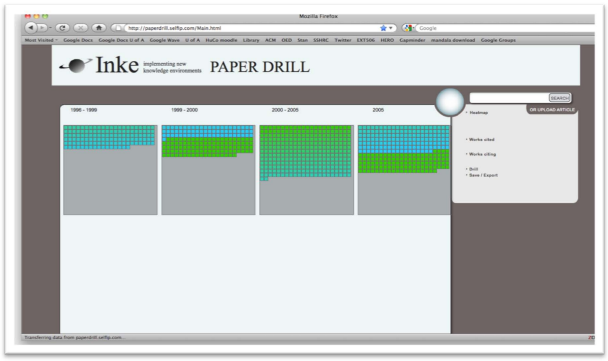


Figure 2: Paper Drill prototype

In the first year of INKE, we began to explore the possible affordances of a tool to support scholars in following citation trails through collections of academic articles (Ruecker et al. 2009). Our approach was to prototype a tool called The Paper Drill, where the user could choose a seed article, then obtain a report about the most frequently cited authors or articles. In year two, we are continuing development of the Paper Drill, but have added the concept of showcase browsing, where the home page of the tool provides heatmaps of the most frequently cited articles, arranged according to date range and journal category. The purpose of this overview is to add information that is not otherwise available in the display until after a report has been generated. We have also been experimenting with data sources, since the system becomes more effective (and the showcase visualization a more complex challenge) as the metadata in humanities journals improves. We will demonstrate The Paper Drill working on a journal article collection.

PAPER 4

Diachronic View on Digital Collections Interfaces

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta

This presentation looks at the interface design of three successful and important digital collections: Project Gutenberg (http://www.gutenberg.org/), Perseus Digital Library (http://www.perseus.tufts.edu/) and the Victorian Web (http://www.victorianweb.org/index.html). Each of these projects has existed for over 10 years and gone through multiple interface designs, sometimes across technologies, which makes them ideal subjects for a diachronic analysis of the evolution of interfaces to digital collections. Studying interface evolution over time helps us understand:

1. The relationship between the nature of the functionalities provided and interface design;
2. The perceived or real differences between different types of projects from open source projects like Project Gutenberg to scholarly editorial projects like the Rossetti Archive;
3. The pace of adopting and integrating new technologies and design perspectives; and
4. The amount and nature of the influence exerted by user-demands over design-decisions.

The methodology of the study includes, but is not limited to, a review of the existing literature about the history, architecture and design of the three collections, an environmental scan of all available versions of interfaces employed, and, where possible, interviews with people involved in the projects.

We will conclude our presentation by acknowledging the role played by interface design in the success of these collections and by assessing the value of the diachronic approach adopted in our study for interface analyses.

PAPER 5

The Corpus from Print to the Web

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

In this presentation, we will examine the design and evolution of corpus interfaces by comparing features from two epochs of information design, the epoch of print and that of the web. In order to survey this variety of design we will present a draft topology of corpus design patterns. Patterns that we have identified include legal, religious, literary and archaeological corpora.

Notable print corpora include the 6th century *Corpus Juris Civilis*, a collection of Roman laws that had as a notable effect the disappearance of the original writings on which it is based. In effect the collection replaces the individuals from history. Another significant example is *The Royal Imperial Coinage*, a print catalogue of Roman coins equipped with a variety of complex and innovative search and access interfaces.

We will then show the creative ways in which corpus interfaces have been adapted to the online environment, along with the attendant novel methods of access and navigation that are then offered. For instance, the *Corpus Inscriptionum Latinarum*, begun in print in 1853 and now being maintained and developed on the web, illustrates the effects of web remediation on searching and layout for traditional corpora. The *Pyramid Texts Online* recreates the original physical structure of the corpus, offering an inventive map-based navigational interface leading to translations and photographs of original texts.

To conclude, this paper will reflect back on the development of the corpus from print to web, and will show how interface features within particular patterns have matured in the move to the web.

PAPER 6

Structured Surfaces for JiTR

Radzikowska, Milena
mradzikowska@gmail.com
Mount Royal University

Ruecker, Stan
sruecker@ualberta.ca
University of Alberta

Brown, Susan
sbrown@uoguelph.ca
University of Guelph

Organisciak, Peter
organis2@illinois.edu,
University of Illinois at Urbana-Champaign

In this last presentation, we describe the usefulness of a structured surface in the design of human-computer interfaces to collections, and propose as an experimental design idea user-generated structured surfaces that can be controlled in an interactive manner. These surfaces are interfaces that structure items in a collection in different ways. A structured surface is a cognitive interface artifact that provides a layer of meaning that supports the data imposed upon it.

Figure 3: Nightingale's Rose Interface

We experimented with a mashup of content provided by JiTR (Rockwell et al. 2009, Rockwell 2008) and a structured surface inspired by Florence Nightingale's rose diagram. The surface is composed of wedges representing months and a series of segments representing cause of death. That's where the original information graphic ends, but we propose adding an additional layer of information. In this case, dots

representing medicine shipments. Alternatively, the dots could represent the individual deaths of soldiers from Shropshire or number of ambulances with flat tires. In the context of JiTR, depending on your metadata, the wedges might represent genre, the segments authorship, and the dots are individual people who were at last year's DH conference. We should mention that the structured surface could be one of a number of preexisting visualizations. Our intention is to experiment with Stanford's excellent collection at Protovis.com (Heer et al. 2010). We feel that the third layer of information represented by the dots or pins provides an exciting opportunity for people assembling dynamic collections to feed into the text analysis tools available through JiTR.

References

Bush, V (1945) (July 1945). 'As We May Think'. *Atlantic Monthly.* 101-08.

Heer, J, Bostock, M, Ogievetsky, V (2010). 'A Tour through the Visualization Zoo'. *Queue.* 5: pp.20-30.

Kirschenbaum, M (2004) (2004). 'So the Colors Cover the Wires: Interface, Aesthetics, and Usability'. *A Companion to Digital Humanities.* Schreibman, S , R. Siemens, J Unsworth (eds.). Oxford: Blackwell. http://www.digitalhumanities.org/companion/.

Hamilton, E, Cairns, H (eds.) (1961). *The Collected Dialogues of Plato including the Letters.* Princeton: Princeton University Press.

Rockwell, G, Stan R, Organisciak, P, Sinclair, S (2009). 'Ubiquitous text analysis'. *Digital Humanities 2009 conference.* University of Maryland.

Rockwell, G (2008). 'Just In Time Research (JiTR): Supporting Experimental Text Analysis.'. *CaSTA 2008, New Directions in Text Analysis* Paper presented at conference. University of Saskatchewan Saskatchewan.

Ruecker, S, Rockwell, S, Radzikowska, M, Sinclair, S, Vandendorpe, C, Siemens, R, Dobson, T, Doll, L, Bieber, M, Eberle-Sinatra, M, Lucky, S, The INKE Group (2009). 'Drilling for Papers in INKE'. *Proceedings of the INKE 2009: Research Foundations for Understanding Books and Reading in the Digital Age*.

Shneiderman, B, Williamson, C, Ahlberg, C (1992). 'Dynamic queries'. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92, the SIGCHI conference.* Monterey, California, pp. 669-670. http://portal.acm.org/citation.cfm?doid=142750.143082.

Shirky, C (2008). 'It's Not Information Overload. It's Filter Failure'. *Web 2.0 Expo NY*Web 2.0 Expo NY viewed 1 November 2010 http://web2expo.blip.tv/file/1277460/.

Papers

Automatic Extraction of Hidden Keywords by Producing "Homophily" within Semantic Networks

Akama, Hiroyuki
akama@dp.hum.titech.ac.jp
Tokyo Institute of Technology

Miyake, Maki
mmiyake@lang.osaka-u.ac.jp
University of Osaka

Jung, Jaeyoung
catherina.rosset@gmail.com
Tokyo Institute of Technology

A complex network is usually conceived of in the form of a graph consisting of nodes representing individual, or atomic, entities and edges linking them according to information about semantic attributes or some weighting value. This intellectual intuition can be applied to the world of language where a sign makes sense through the concurrent presence of other signs (Saussure, 1916). In particular, the semantic aspect of a document, as a group of more or less coherent sentences, depends on co-occurrence information about the words being bound together to form topics. When a content-bearing word is found within a neighborhood of other words, one may assume that they produce a constellation of meaning (Schütze, 1997; Takayama, 1999). Conceptual interrelatedness can be represented in a graph form called a semantic network. Similarly, in the field of digital humanities, the graph-based approach to document analysis is becoming a major research trend (Miyake, 2009).

In semantic networks, graph coefficients are useful for examining the features of language data, such as a document or corpus, from the perspective of meaning. Hubs with the highest degree values can be regarded as being as key words (excluding functional words), that are highly involved in producing the document context. In normal cases, the bias of degree distributions to follow a power law (or, more concretely, Zipf's law) has been handled by the concept of scale-free (Barabási et al., 1999) for complex networks. However, one graph index that researchers are increasingly paying close attention to is the concept of 'intrinsic weights' that are not distributed to edges but to vertices (Caldarelli et al., 2002; Boguñá et al., 2003; Masuda et al, 2004). The emergence of this weighting within some network settings certainly leads a situation where vertex degree generates a phenomena known as the 'Rich Club' (Zhou et al, 2004; Colizza et al., 2006), consisting of hubs with high intrinsic weights. However Masuda et al. (2006) have revealed a contrastive kind of subgraph area, which they figuratively refer to as a 'VIP Club', where vertices with similar weighting values are exclusively connected to form a confined circle of privileged elites. This "homophily" tendency (Axelrod, 1997; Barrat et al., 2004; Centola et al.2007) makes the average shortest path length significantly longer, so that the innermost graph area becomes harder to access, forming a preserve for the so-called 'masterminds' (Masuda et al., 2006).

These elite entities may be, in some sense, regarded as 'hidden keywords' within complex networks, where all the words are interconnected—at dense intervals—from a paradigmatic perspective (Jacobson, 1963). These mastermind words are unquestionably of moderately high frequency or degree, but they are tightly related to one another by 'homophilous' ties to form very important but discreet lexical patterns. The term homophilous here means a significantly high value of 'degree correlation' (Boguñá et al, 2003; Boccaletti et al., 2006), which serves as a marker for the VIP-Club phenomenon (See Figure I). These mastermind words, which are relatively difficult to retrieve during comprehension within the reading process, sometimes play a crucial role in producing well-calculated subliminal effects or hinting at authors' obsessions with their long-cherished themes.

Be that as it may, the Incremental Advancing Window (IAW), a windowing method (Burges, 1998; Lemaire et al., 2005) that Akama et al (2008) have proposed in order to extract word association patterns from a whole document, clearly satisfies the requirements for creating a homophilous semantic network from lexical co-occurrence information. From such a graph, it is possible to extract hidden keywords with moderate frequencies as members of the clusters regarded as being VIP-Clubs. In this method, the window proceeds step by step through an entire document (after the removal of noise words) to collect all word pairs appearing at least once inside the frame. The list of all pair instances thus obtained with their frequencies makes it possible to generate a semantic network. However, two parameters are set to some specific values to adjust for recall and precision in the data gathering: namely, the window size (diameter) is changed from 1 to m and the threshold for word pair frequency (theta; θ) is changed from 1 to n (m and n

are both natural numbers larger than 1). For example, if the theta value is 3, word pairs appearing less than 3 times in the window are ignored. Precisely, this means that, no matter how frequently a word appears in a text, keywords are rejected from the list of vertices for the semantic network representing the text, if the instances of paired words are extremely rare with recorded frequencies lower than θ. This is why a graph derived by IAW exhibits the homophily tendency, if we consider degree itself to be an intrinsic weighting for a vertex.

For instance, let us cite a study conducted by Akama et al (2008) which applied IAW to Saint-Exupéry's novel "Le petit prince" (original French version). The sample consisted of 1,312 content-bearing words remaining after a stop list was applied. If window size (diameter) is at the smallest level and threshold is similarly maximally strict, then the words extracted by IAW are numerically low, but they form tightly cohesive aggregates, suggesting that the precision rate, P, and the recall rate, R, are always in a trade-off relationship. A severe windowing condition with a threshold value set to 6 allowed us to recall the 38 most important words from the standpoint of co-occurrence patterns, but some of them were not included in the list of the 38 most frequent words. These hidden keywords, or mastermind words, were «apprivoiser», «boa», «cent un», «consigne», «fermé», «manger», «monde», «posséder», «ramoner», «région», «unique», and «volcan», which could all be characterized as homophilous in terms of degree similarity.

In contrast, words that were among the top words but excluded from the most severe IAW computational conditions were «aimer», «ami», «baobab», «connaître», «croire», «dessin», «jour», «nuit», «grande personne», «rose», «sérieux», and «venir», which are all ordinary keywords sharing the feature of heterophily in terms of degree-frequency —a tendency to be linked, or even collocated, with many different words (Figure II). In the smallest semantic network containing both groups of words (window size = 7, threshold for pair frequency = 6), the average shortest path lengths from all dangling nodes representing peripheral words to the vertices of these 12 masterminds (the first group) and to those of the purely general keywords (the second group) were 4.172 and 3.747, respectively, representing a highly significant difference according to a conducted t test (Figure III).

This result reveals much about the characteristic traits of mastermind words, which when coupled with graph clustering, permits us to understand how they shape communities that deserve the label of VIP-Clubs. To prove this, Markov Clustering (MCL) was applied to the smallest semantic network with IAW parameters of window size = 1 and threshold for pair frequency = 6 (thus screening out many weak co-occurrence patterns and focusing on the most frequent instances of bi-gram). MCL proposed by Van Dongen (2001) is well known as a scalable unsupervised cluster algorithm for graphs that decomposes a whole graph into small coherent groups by simulating the probability movements of a random walker across the graph. It is assumed that when MCL is applied to semantic networks, it yields clusters of words that share certain similarities in meaning or that appear to be related to common concepts (Dorow et al., 2005; Jung et al, 2007). As a matter of fact, the present clustering results produced some clusters which consist almost exclusively of homophilous masterminds, such as {monde, unique}, {boa, fermé, serpent}, and {éteindre, volcan, ramoner}. These clusters suggest a series of subtopics that are not so dominating within the novel, but that convey a persistent and deep resonance to the readers.

As described, a text, when treated in the form of a graph, exhibits some hidden keywords that can be enumerated as mastermind entities through the analysis of a homophilous semantic network. Furthermore, if a graph clustering method, such as MCL, is applied to the network, the vertices with such features are categorized into sub-topic clusters, known as VIP-Clubs. Despite the moderate degrees (frequencies) of these vertices (words), they are inconspicuously combined to create lexical patterns which, although they are minor, or subsidiary, in nature, are yet effective.

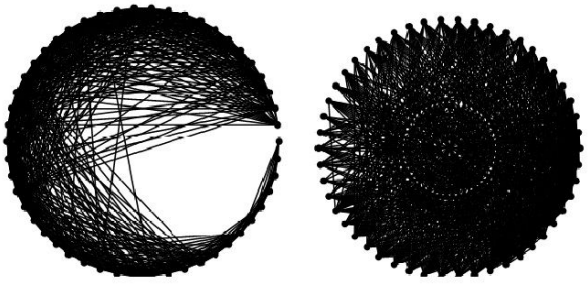


Figure I : A homophilous graph (left) and a heterophilous graph (right) : an agglomeration of 'VIP-Club' is recognizable in the homophilous graph, while homogeneity underlies the heterophilous graph. Both networks with the same number of nodes (50) and the same connectivity (38.5%) are produced by respectively applying different *pruning functions*-- which consist of trimming edges with probabilities

varying according to the weight correlation--to an identical random graph whose distribution of 'intrinsic weights' follows the degree distribution of the equal-sized BA model (scale-free graph). The two pruning functions are

$$P_{homo} = 1 - \frac{1}{1 + Exp[(-c)(x - Median[abs\log diff])]} \quad P_{hetero}(x) = \frac{1}{1 + Exp[(-c)(x - Median[abs\log diff])]},$$

where *abs*log*diff* stands for the absolute value of the logarithmically-transformed difference of 'intrinsic weights' between any two vertices.

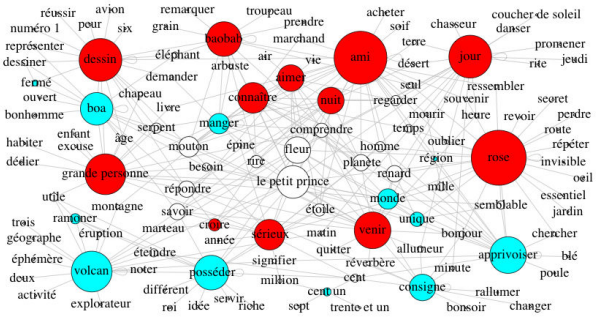


Figure II: Subgraph (extracted from the semantic network made under the condition of window size = 7, threshold for pair frequency = 6) around the homophilous words (masterminds: blue) and the heterophilous words (purely general keywords: red). The size of a vertex corresponds to the degree.

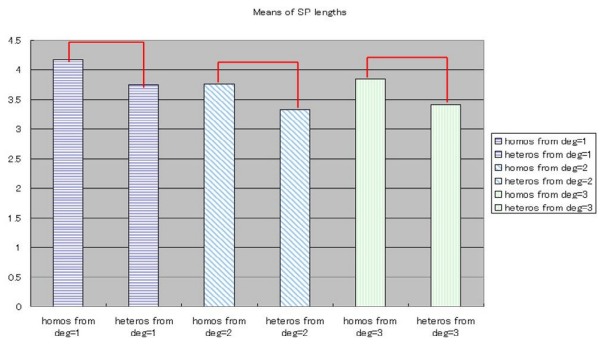


Figure III : The average shortest path lengths from the words with degrees 1, 2 and 3 respectively to the homophilous words (masterminds) and to the heterophilous words (purely general keywords)

References

Akama, H., Miyake, M., Jung, H. (2008). 'A New Evaluation Method for Graph Clustering of Semantic Networks Built from Lexical Co-occurrence Information'. *Post-Conference Journal Paper of the 18th International Congress of Linguistics (CDROM)*.

Axelrod, R. (1997). 'The Dissemination of Culture: A Model with Local Convergence and Global Polarization'. *Journal of Conflict Resolution.* 2: 203-226.

Barabási, A.-L., Albert, R. (1999). 'Emergence of scaling in random net-works.'. *Science.* 286: 509-512.

Barrat, A., Barthelemy M., Pastor-Satorras, R., Vespignani, A. (2004). 'The architecture of complex weigthed networks.'. *Proc. Natl. Acad. Sci. USA..* 11: 3747-3752.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang D.-U. (2006). 'Complex networks: Structure and dynamics'. *Physics Reports.* 4-5: 175-308.

Boguñá, M., Pastor-Satorras, R., Vespignani A. (Jan 20, 2003). 'Epidemic spreading in complex networks with degree correlations'. *cond-mat.stat-mech.* .

Burgess, C., Livesay, K., Lund, K. (1998). 'Explorations in context space: words sentences and discourse.'. *Discourse Process.* 211-257.

Caldarelli, G., Capocci, A., De Los Rios, P., Munoz, M.-A. (2002). 'Scale-free networks from varying vertex intrinsic fitness'. *Physical Review Letters.* 258702.

Centola D, González-Avella J.-C., Eguíluz V.-M., San Miguel, M. (2007). 'Homophily, Cultural Drift, and the Co-Evolution of Cultural Groups.'. *Journal of Conflict Resolution.* 6: 905-929.

Colizza, V., Flammini, A., Serrano, M.-A., and Vespignani, A. (2006). 'Detecting rich-club ordering in complex networks'. *Nature Physics.* 2: 110-115.

Dorow, B., Widdows, D., Ling, K., Eckmann, J.-P., Sergi, D., Moses, E. (2005). 'Using Curvature and Markov Clustering in Graphs for Lexi-cal Acquisition and Word Sense Discrimination.'. *MEANING-2005, 2nd Workshop or-ganized by the MEANING Project.* February 3rd-4th, 2005..

Jakobson, R. (1963). *Linguistique et théorie de la communication: Essais de linguistique Générale..* Paris: Les Éditions de Minuit.

Jung, J., Akama, H. (2008). 'Employing Latent Adjacency for Appropriate Clustering of Semantic Networks'. *New Trends in Psychometrics.* Tokyo: Universal Academy Press, pp. 131-140.

Lemaire, B., Denhière,G. (2004). 'Incremental Construction of an associated Network from a Corpus.'. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society.* Pp. 825-830.

Masuda, N., Miwa, H., Konno, N. (2004). 'Analysis of scale-free networks based on a threshold graph with intrinsic vertex weights'. *Phys. Rev. E.* 70: 036124.

Masuda, N., Konno, N. (2006). 'VIP-club phenomenon: Emergence of elites and masterminds in social networks'. *Social Networks.* Volume 28, Issue 4: 297-309.

Miyake, M. (2008). 'Investigating word co-occurrence selection with extracted sub-networks of the Gospels Employing Clustering Coefficients'. *Digital Humanities.* VENUE, 2008, pp. 258-260.

Saint-Exupéry, A. de. (1971). *Le Petit Prince.* Harcourt, Brace & World, Inc..

Saussure, F. de. (1916). *Cours de linguistique générale.* Payot.

Schütze H., Pederson, J.-O. (1997). 'A coocurrence-based thesaurus and two applications to information retrieval'. *Information Processing & management.* 33(3): 307-318.

Takayama Y. et al. (1999). 'Information Retrieval Based on Domain-Specific Word Associations'. *Proceedings of PACLING '99.* Waterloo, Ontario, Canada, June 1999.

Van Dongen, S. (2000). *Graph Clustering by Flow Simulation.* Amsterdam: University of Utrecht PhD thesis.

Zhou, S., Mondragon, R.J. (2004). 'The rich-club phenomenon in the Internet topology'. *IEEE Communications Letters.* 180-182.

The Text-Image-Link-Editor: A tool for Linking Facsimiles & Transcriptions and Image Annotations

Al-Hajj, Yahya Ahmed Ali
alhajj@fh-worms.de
Worms University of Applied Sciences, Germany

Küster, Marc Wilhelm
kuester@fh-worms.de
Worms University of Applied Sciences, Germany

1. Introduction

TextGrid's Text-Image-Link-Editor (TBLE "stands for for Text-Bild-Link-Editor, German for Text-Image-Link-Editor") is used to link segments of text with sections on the corresponding image. A typical application is linking of scans of facsimiles with their transcriptions, though these texts can also be created during the linking process, which allows e.g. also for image annotations. The information on the linking between manuscript fragments and the corresponding transcription is itself stored in TEI. TextGrid is as a virtual research environment (VRE) for the humanities disciplines dealing with texts in a wide sense (philologies, epigraphy, linguistics, musicology, art history etc.). The joint research project TextGrid is part of the D-Grid initiative, and is funded by the German Federal Ministry of Education and Research (BMBF) for the period starting June 1, 2009 to May 31, 2012 (reference number: 01UG0901A).

TextGrid consists of two principal building blocks, the grid-based backend TextGridRep that hosts both infrastructure services and the repository layer for access to research data and longterm archiving, and the user-facing TextGridLab. The TextGrid Laboratory (TextGridLab), a single point of entry to the virtual research environment, will provide integrated access to both new and existing tools and services via a user friendly software [TG]. TBLE is a key component of the TextGridLab that has been under continuous development since 2008 and is by now in practical use.

2. State of the Art

The integration of manuscript scans with their transcription and indeed the critical edition itself is a

desideratum of modern editions: “While some people continue to think of electronic texts as exclusive of images, the fact is that digital images of manuscripts are electronic texts, as well. The most compelling scholarly editions of the future will make full use of markup schemes such as XML [...], but not without extensive integration of images” [Kiernan2006]. In this context TextGrid is not the only project that recognized the need for an tool to facilitate this linking of image sections with transcriptions. The Edition Production & Presentation Technology's (EPPT's) [EPPT] tool box for integrating images and text operates in much the same solution space. [Parker2009] proposes the development of a web-based Text-Image Linking Environment (TILE), and for much the same reasons as the TBLE, namely to facilitate “the linking of images and textual information [which] remains a slow and frustrating process for editors and curators”. [TILE] is currently under development.[1] Unlike the Eclipse-based TBLE, TILE is Ajax-based, extending the Ajax XML Encoder.

Similar in objective to TBLE and developed in much the same timeframe is Tapor’s / the University of Victoria’s Image Markup Tool [Holmes2010]. Both independently decided to use formats based on TEI P5 to store linking information, though at this stage unfortunately not the same. Unlike TBLE, the Image Markup Tool is a desktop program only for the MS Windows platform and cannot be integrated into the TextGridLab.

[Cayless2008] reports on experiments to partially automate the linking between manuscripts and their transcriptions. TBLE plans to integrate similar functionality using OCR technology (cf. below).

As required in [Huitfeld2010], TBLE allows, in Peirce’s terminology, multiple “types” to be associated with one “token”, or in other words to associate one section in a manuscript with multiple, possibly contradictory interpretations / transcriptions. Image sections can overlap, so that divergent segmentations are possible.

As an aside, this type of linking is very different from the research field of automated image analysis and image annotation which attempts to automatically establish key metadata for an image, e.g. by identifying the objects or persons shown on a photo.

3. Functionality

The following use case is a typical example for working with the TBLE: A scholar wants to publish manuscripts as collection of images, which offer a digital representation of the original work, but also wants to publish his take on its correct transcription in view of establishing a critical edition. The solution is to write the content of these hand written documents as text in a human/machine readable format e.g. XML and this text can be divided into logical related segments for example: verses, lines and then these text segments can be easily linked with the corresponding sections on the images using the TBLE.

The TBLE can be used for:

3.1. Linking Existing Texts and Digitized Manuscripts

Text and image are opened, then the corresponding components (text segment and selected image section) are marked by pairs and the linkage is confirmed. The results can be saved as a new file (local or in the TextGridRepository), which contains the linking information (image coordinates, text segment identifiers, URIs of the used text and image files). Sections can be rectangles or arbitrary polygons.

The content of the new created file represents the saved information as a TEI document with an embedded [SVG] section (see section 4 “The TEI-Model” for more details). Once a file is saved, double clicking it reloads used images, texts and links to continue editing. Changing or adding new links as well as modifying the linked text is possible at any time.

Instead of starting out with an existing transcription and linking it with the image data, the scholar can also decide to start from scratch with an empty text file. The new text segments can be inserted stepwise or at once.

Any number of different and possibly conflicting transcriptions and segmentations can be linked against one set of digitized manuscripts.

3.2. Annotation of Image Sections and Existing Links

The Text-Image-Link-Editor offers many other useful features, that help annotating specific links or image sections. For example:

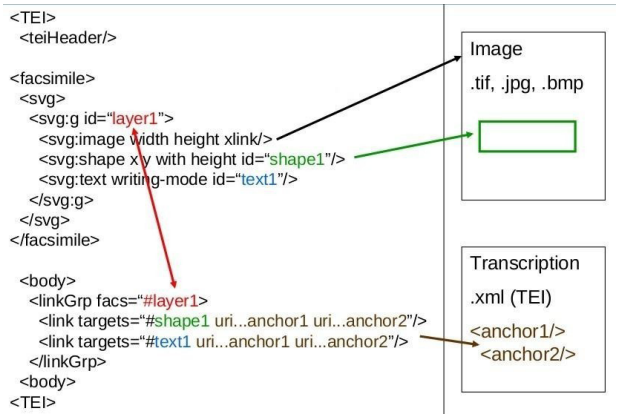
1. It's possible to build logical groups of links (e.g. verses, comments, etc.) using the layers-tool.
2. A text-direction (e.g. left-to-right & top-to-bottom) can be assigned to the links.

4. The TEI-Model

The output file of the Text-Image-Link-Editor follows the TEI model with embedded svg description

elements. The following is a list, which crudely describes the structure of the TEI document:

1. **<teiHeader>** this element could be used to save the metadata of the document.
2. **<facsimile>** in this element is the svg element embedded, which keeps the topographic descriptions of images and links.
3. **<body>** in the body element are the link groups, that contain the link elements. These link elements represent the relationship between the image sections and the corresponding text segments. The relationship between images and texts and links is represented in the following figure:



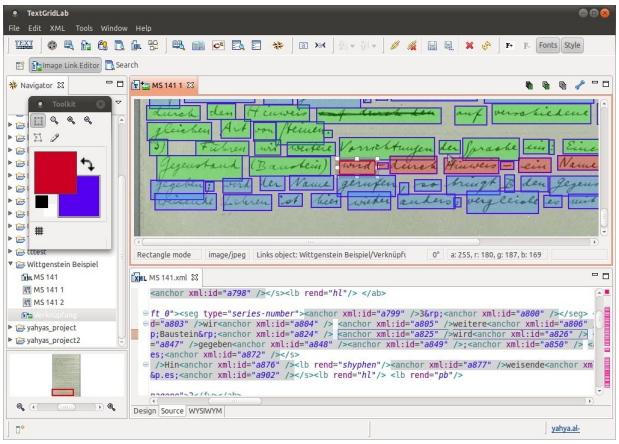
The TEI output file of the Text-Image-Link-Editor to describe the relationship between images and texts and links [Maynooth2010]

5. Structure and Components

The Text-image-Link-Editor is a part of the TextGridLab and is implemented as a group of Eclipse plugins following the [MVC] pattern[2].

This tool consists of a Toolkit and two views in addition to the XML Editor and the generic Navigator:

- **Image View:** shows the images and enables you to select individual image sections to be linked.
- **Thumb View:** is used for navigation. It displays a reduced version of the entire image and the active image detail (which is enlarged in the Image View) which can easily be moved and zoomed.
- **Toolkit:** provides different functions for working on the Image View.
- **XML Editor:** allows you to open or create texts as well as to select individual text elements.



TBLE in action

6. Further Enhancement

TBLE is already actively used in a number of projects, but continues to be enhanced, taking into account new user requirements coming up in the field. In particular, we plan to implement a new feature in the Text-Image-Link-Editor to enable an automated segmentation of facsimiles using the [OCRopus] OCR-system, which offers a possibility to partially automate the linkage process.[3]

References

Cayless, Hugh (2008). 'Experiments in Automated Linking of TEI Transcripts to Manuscript Images'. *TEI Members Meeting.* 2008`http://www.cch.kcl.ac.uk/cocoon/tei2008/programme/abstracts/abstract-166.html`.

Kiernan, Kevin et al.. *Edition Production & Presentation Technology (EPPT).* `http://www.eppt.org/eppt/` (accessed 2011-03-12).

Holmes, Martin. *The UvC Image Markup Tool Project. Project homepage.* `http://tapor.uvic.ca/~mholmes/image_markup/index.php` (accessed 2010-10-31).

Huitfeldt, Claus; Yves Marcoux and C. M. Sperberg-McQueen (2010). 'Extension of the type/token distinction to document structure'. *Proceedings of Balisage: The Markup Conference 2010.* Balisage Series on Markup Technologies.

Kiernan, Kevin (2006). 'Electronic Textual Editing: Digital Facsimiles in Editing'. *Electronic Textual Editing.* Burnard, Lou; O'Brien O'Keeffe, Katherine and Unsworth, John (Eds) (ed.). MLA. `http://www.tei-c.org/About/Archive_new/ETE/Preview/kiernan.xml` (accessed 2011-03-14).

Maynooth - Michael Leuk, Dr. Simon Rettelbach (2010). *Cost Workshop*.

Porter, Dorothy Carr; Reside, Duke andWalsh, John (2009). 'Text-Image Linking Environment (TILE)'. *Digital Humanities, 2009.* 2009, pp. p. 388ff.

W3C, Scalable Vector Graphics (SVG). `http://www.w3.org/Graphics/SVG/` (accessed 2011-03-14).

The Text Encoding Initiative. `http://www.tei-c.org` (accessed 2011-03-14).

About TextGrid. `http://www.textgrid.de/en/ueber-textgrid.html` (accessed 2011-03-14).

Notes

1. For the current status of the project cf. also its homepage `http://tileproject.org` Consulted 2011-03-14
2. The Model-View-Controller (MVC) pattern separates the modeling of the domain, the presentation, and the actions based on user input into three separate classes
3. OCRopus (tm) is a state-of-the-art document analysis and OCR system, featuring pluggable layout analysis, pluggable character recognition, statistical natural language modeling, and multi-lingual capabilities. http://code.google.com/p/ocropus/ . Consulted 2011-03-14
4. doi:10.4242/BalisageVol5.Huitfeldt01.

Content Patterns in Digital Humanities: a Framework for Sustainability and Reuse of Digital Resources

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

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

Research in the arts and humanities has created much digital material that represents a significant investment, both of funding and of intellectual effort. In the UK at least, given the current lack of national infrastructure for sustaining this material, these resources are typically hosted in their home institutions using a variety of approaches and technologies.

This incurs a number of risks. At the most basic level, without ongoing maintenance a resource ceases to be usable at all as the technologies in which it was implemented become obsolete and unsupported. Even if hosting institutions apply preservation techniques to ensure continued accessibility of resources, this does not enable collections to make full use of technological advances that might greatly enhance their utility for and impact on research. Access to legacy resources may be limited to a simple download or browser access in a website. In neither case does this facilitate advanced research services, such as mashups or data/text mining, that will become increasingly common in future digital research.

The impact of humanities research may still be felt many years after the original research was undertaken – the information produced has a long lifespan in intellectual terms. Sustainability does not just mean keeping the data alive, but enabling the exploitation of advances both in technology – making the data accessible in new ways – and in humanities research – forging connections between resources that lead to new discoveries and broader impact.

Digital resources often exist in "silos", lacking interoperability. Individual projects typically address focused topics, and may implement digital resources in idiosyncratic ways and to address their immediate needs. This results in a multitude of resources that

are scattered and disparate in nature, yet related intellectually, resources that, linked up, would form a whole much more useful for research than the sum of the parts, much as fragments of a map, when combined, allow navigation from one place to another. Ultimately, the vision here is of a virtual and distributed "web of knowledge".

The digital resources in the humanities may be characterised by their diversity and complexity. Collections involve multiple media and standards. The material may be highly complex, with many structural and semantic relationships both internal and contextual; the interpretation of an object (e.g. an inscription) may depend on its relationships to other resources (e.g. other inscriptions/texts, surveys, concordances).

One approach to this would be to develop enhancements to individual resources; however, to be truly sustainable we should avoid such ad hoc solutions. The primary question asked by our project is thus – how can we develop a generic framework for digital resources in the arts and humanities that addresses the above issues for a broad range of collections, and that is not a closed system but can be extended to support other digital material and (possibly unanticipated) future tools, technologies and research methods?

We are attempting to answer this question in the CMES (Content Models for Enhancement and Sustainability) project, which is funded by the UK Arts and Humanities Research Council as part of its DEDEFI (Digital Equipment and Database Enhancement for Impact) programme. We are developing a framework using the Fedora digital repository software for sustaining and enhancing particular groups of digital resources produced by earlier digital humanities. We are addressing two groups of collections, each typical of a wide range of humanities research activities:

1. Digital texts, which may comprise complex networks of diverse information: images, markedup text, geospatial data, translations, standoff annotations/ markup, and potentially extensive links to external resources. We are addressing two groups of resources managed at KCL: the Stormont Papers, and the Inscriptions of Aphrodisias. These contrasting examples – one modern and dealing with large, complete volumes, one ancient and dealing with small, fragmentary texts –facilitate development and testing of generic models. They also provide scope for demonstrating the utility of our framework for (i) developing new material (e.g. Stormont parliamentary papers and Inscriptions of Roman Tripolitania/Cyrenaica), and (ii) forging links with external digital material (e.g. Westminster Hansards and Pleiades).
2. Multimedia performing arts collections, specifically the following resources managed at KCL: Scottish Traditions of Dance, which contains text, images, video, interviews, audio and databases, Adolphe Appia, which contains images, 3D virtual reality models, and audio from the King's Sound Archive.

Fedora is particularly good for modelling complex material and links between objects. Representations of digital objects within Fedora are formalised as "content models" (henceforth, CMs), which may be regarded as "data types" for digital objects. We will review the selected collection groups and develop a set of CMs that support them by providing consistent, standardised and interoperable (yet flexible) patterns for representing these collection types. We will need to go beyond Fedora's relatively simple CM formalisation to produce these "Content Patterns" (henceforth, CPs) for complex collections, e.g. by using the Enhanced CM framework developed by SULD, which allows the specification of relationships and ontologies, and the definition of collection templates.

We analysed the resources along with subject specialists in digital text and performing arts resources. Note that, given the variation in how legacy collections have been implemented, the CPs may be idealisations that do not directly match the collections, which may require a degree of reworking to make them fit. We will not be overprescriptive here – diversity arises naturally from the research material – but a degree of common practice would be beneficial for the creation and reuse of the material. Moreover, our CPs will provide foundations that can be extended easily to support diverse community practices.

Each of the target collections had its own custom web interface, driven by quite different underlying data models. We are developing consistent delivery/ publishing mechanisms for the different collection groups that are driven by the underlying CPs. This has the benefit that these mechanisms are available for any collection that conforms to the CP, leading to more consistent and interoperable interfaces for resources of similar type.

However, this will not necessarily lead to homogeneity. Our approach enables the structure of collections to be represented with fine granularity, and interfaces are correspondingly modular. This facilitates the creation of more integrated web views across different collections, but it also allows content to be exposed as machinereadable feeds that can be

used to provide addedvalue services, e.g. aggregating content, automated processing (e.g. text mining), mashups etc. The creators (or curators) of a resource will no longer be the arbiters of how information should be delivered and used. The resources produced by research are not just ends in themselves – they provide source material for subsequent research – and to maximise impact they should be made available in ways that allow scholars unrelated to the original editors to make transformative use of them, rather than just via a website. We are thus providing a framework whereby users (perhaps domain experts) can develop and integrate their own tools to process resources.

The project is thus not only enhancing particular collections, but producing a framework that is extensible in several ways:

- The generic CPs and associated tools provide templates for simplifying creation of new collections of similar form (e.g. digital texts), and guarantee certain functionality that conformant collections would inherit from the template.
- The set of content patterns is itself extensible, following the same methodology, to other collection types.
- The framework can be extended with new tools as technologies change (tools/services can be linked to CPs and inherited by collections that follow the pattern).

The project thus builds on existing efforts and provides a foundation for a broader and longerterm programme for sustaining and enhancing digital humanities research. Developing this framework to support resources based around digital texts and performing arts will cover a significant amount of ground, and provide a springboard for future extensions. It will also ensure sustainability by integrating these initiatives into repository and curation infrastructures at KCL, and will allow a growing corpus of digital material to be integrated into this infrastructure.

Enroller: A Grid-based Research Platform for English and Scots Language

Anderson, Jean
Jean.Anderson@glasgow.ac.uk
University of Glasgow

Alexander, Marc
Marc.Alexander@glasgow.ac.uk
University of Glasgow

Green, Johanna
Johanna.Green@glasgow.ac.uk
University of Glasgow

Sarwar, Muhammad
Muhammad.Sarwar@glasgow.ac.uk
National e-Science Centre, UK

Sinnott, Richard
rsinnott@unimelb.edu.au
University of Melbourne

1. Summary

This paper describes a collaboration between eScientists and humanists; specifically Grid scientists and language and literature scholars, working together to create a repository of data sets and tools, combining our most advanced knowledge in all areas.

Language and literature scholars make use of variety of language resources to conduct their research. Such resources include dictionaries, thesauri, corpora, images, audio and video collections. At present most of these resources are distributed, non-interoperable and license protected. As a result researchers typically conduct their research through direct access to independent data sets using multiple browser windows and multiple authorisations. This approach results in non-scalable and less productive research, and often leads to incomplete and/or non-verifiable results.

The JISC funded project, *Enhancing Repositories for Language and Literature Researchers* (Enroller) is addressing these issues through development of a targeted eResearch environment. This paper presents the current state of progress and outlines how secure access to distributed data resources with targeted analysis and collaboration tools is supported. In the full paper we will also describe how Enroller is exploiting

high performance computing infrastructures such as the UK National Grid Service and ScotGrid, and discuss a problematic issue that has arisen through the differing working practices of humanists and scientists.Consider a typical language and literature scenario where a researcher wants to search for a word such as "bonny" in the dictionary to find its meaning; in a thesaurus to look up the concepts and categories it is found in and in a corpus to find the documents containing it. The user may also want to see the concordances and word frequency of the word in each found document. In undertaking this process, they might want to save the different results; share them with others and possibly perform a comparison between many different resultant data sets. This scenario becomes more challenging when multiple dictionaries, thesauri and text corpora need to be cross-searched simultaneously, for example when the researcher wants to lookup the word "bonny" in the Oxford English Dictionary, in the *Scottish National Dictionary*, and in the *Dictionary of the Older Scottish Tongue*. The researcher may also want to search for the word in the *Historical Thesaurus of English* to look up synonyms and related concepts and categories and then search for all of the matching concepts in further corpus resources.

Researchers will want to use the standard text analysis tools: e.g. concordances, word frequencies, collocation clouds. They may well wish to save and download the results for further analysis or use targeted tools to investigate, e.g. variant spellings of the word 'bonny'. The problem is further exacerbated if the researcher decides to search for multiple, possibly hundreds, of words at once and do all of the mentioned tasks at once. Most of the language and literature data environments do not permit scholars to do any of these activities, instead researchers are left with individual level data sets, coded differently (e.g. with different metadata and data formatting), accessible through individual web-based resources with individual access codes and methods.

The challenge for the project is to maintain the data integrity of its collaborators and the security of access-limited data, while facilitating research across and between each dataset for the benefit of researchers in multiple fields. Enroller provides an interactive research infrastructure that provides seamless, secure access to all of the different language and literature data sets in a user-oriented environment. Furthermore, since many of the searching and analysis efforts can be computationally intensive – especially when bulk searching or building of indexes is required - we provide secure access to high performance computing infrastructures such as the UK e-Science National Grid Service (http://www.ngs.ac.uk) and ScotGrid (http://www.scotgrid.ac.uk). In this project, language and literature researchers, including an associated network of international scholars, are now able to access large amounts of language and literature data and analysis services from a single, easy-to-use portal. Enroller is currently working with the following data sets:

- The EPSRC and AHRC funded *Scottish Corpus of Text and Speech* is a collection of text and audio files covering a period from 1945 to the present. The SCOTS corpus is available in TEI-compliant XML. (http://www.scottishcorpus.ac.uk)
- The AHRC funded *Corpus of Modern Scottish Writing* offers a collection of texts and manuscript images from the period 1700 to 1945. (http://www.scottishcorpus.ac.uk/cmsw/)
- The AHRC funded *Newcastle Electronic Corpus of Tyneside English* provides a corpus of dialect speech from Tyneside in Northeast England. The NECTE corpus is encoded in TEI-compliant XML. (http://www.ncl.ac.uk/necte)
- The *Dictionary of the Scots Language* encompasses two major Scottish language dictionaries: the *Scottish National Dictionary* and the *Dictionary of the Older Scottish Tongue*. DSL data is available in XML format. (http://www.dsl.ac.uk/dsl)
- The *Historical Thesaurus of English* contains more than 750,000 words from Old English to the present in 250,000 categories. HTE was published in print form by Oxford University Press in 2009 and is a new and significant development for historical language studies. (http://libra.englang.arts.gla.ac.uk/historicalthesaurus/)
- The *Oxford English Dictionary* is a commercial resource published by Oxford University Press and is the authority on English language vocabulary. It is accessible through our search interface by API. (http://www.oed.com)

The inclusion of other data sets is underway, e.g. we have incorporated Hansard, early C19th to late C20th, and are negotiating for the 1755 *Dictionary* of Samuel Johnson. Many scholars have no platform or assistance to put texts online and make them accessible to others, far less can they make them interoperable with other relevant data sets. Enroller provides a place where users can deposit their own texts. Texts are wrapped in a basic, TEI minimal XML envelope and indexed. The user can choose whether a text is available to all or private. Once deposited a

text can be searched from the portal with the other data sets.

The project has data at its heart, but it is also exploiting state of the art high-performance computing and security solutions. In particular the project is employing a Virtual Organisation Membership Service (VOMS)-based solution in accessing the NGS where pooled Enroller accounts are used by researchers accessing these resources through a targeted project portal. This includes use and exploitation of the Workload Management System (WMS) to provide resource broking based job scheduling across all of the NGS nodes. This job scheduling is targeted currently to supporting large-scale searching based upon the Google MapReduce application.

The full paper will describe the Enroller project in more detail and outline the capabilities that are supported and our experiences so far in implementing this infrastructure. This includes a description of how the portal has been made accessible through the UK Access Management Federation; the easy to use search system and the reasons for its human-computer interface choices; the advanced Grid-based information retrieval system, capable of executing linguistic workflows, taking advantage of HPC facilities such as NGS and ScotGrid; how the system supports large-scale data indexes for fast searching; support for tools for linguistic analysis, including concordance, collocation and word frequency analysis; support for seamless secure access to licensed data; support for data deposition and automated indexing services; documented analysis of our user experiences in using of the infrastructure provided thus far. We will outline the plans for future adoption by the wider research community and end with our reflections on this eScience/humanities collaboration.

References

J. Watt, .O. Sinnott, T. Doherty, J. Jiang (2008). 'Portal-based Access to Advanced Security Infrastructures'. *UK e-Science All Hands Meeting .* Edinburgh, September 2008, pp. 17.

M.S. Sarwar, R.O. Sinnott (2010). 'Towards a Virtual Research Environment for Language and Literature Researchers'. *IEEE e-Science 2010 .* Brisbane, Australia, December 2010, pp. 8.

Java Database Connectivity (JDBC) API. `http://www.oracle.com/technetwork/java/javase/tech/index-jsp-136101.html` (accessed 15/03/2011).

The Internet2 Shibboleth framework . `http://shibboleth.internet2.edu` (accessed 15/03/2011).

Enroller. `http://www.glasgow.ac.uk/enroller/` (accessed 15/03/2011).

Oxford English Dictionary . `http://www.oed.com` (accessed 15/03/2011).

Scotish Language Dictionaries . `http://www.dsl.ac.uk` (accessed 15/03/2011).

The Scottish Corpus of Texts and Speech. `http://www.scottishcorpus.ac.uk/` (accessed 15/03/2011).

Newcastle Electronic Corpus of Tyneside English. `http://research.ncl.ac.uk/necte/` (accessed 15/03/2011).

The Historial Thesaurus of English . `http://libra.englang.arts.gla.ac.uk/WebThesHTML/homepage.html` (accessed 15/03/2011).

Handling Glyph Variants: Issues and Developments

Anderson, Deborah
dwanders@sonic.net
Department of Linguistics, UC Berkeley

1. Introduction

The challenge of how to handle glyph variants when encoding text has long been a dilemma for those working with historical text materials. How can a digital humanist specify a particular glyph of a Unicode character, even if the glyph might be known to be an error? Is it possible to search for the character, and find instances of the “error” glyphs?

This short paper addresses issues involving glyph variants, in light of recent developments within the world of Unicode and W3C standardization, as well as OpenType specifications (also an ISO standard). The different options for handling glyph variants will also be explored in view of sustainability, and viewed from the general perspective of the Unicode character encoding standard, with particular discussion of Unicode variation sequences.

2. Gaiji

One option available to text encoders is to use the “gaiji” module, a mark-up mechanism described in TEI P5 which offers a means to represent and distinguish specific characters and glyphs that the Unicode Standard considers as identical (TEI Consortium, 2010).

3. Font-Based Option

Another possibility is to request users view the text with a particular font, one that contains the appropriate shape of the glyph(s). However, this is dependent upon the user having a particular font installed.

Two recent developments affect this option:

1. A working group of W3C is developing a specification for “WebFonts,” which will enable the automatic downloading and temporary installment of fonts over the Web, so users don’t need to install fonts on their operating systems. WebFonts is expected to be more widely deployed in the future; a public working draft was published in late July 2010 (W3C, 2010). This would apply to viewing text on Web browsers, and does not currently extend to word processing documents. (Note: W3C also is refining the CSS3 fonts module.)
2. A second development is the OpenType (OT) specification, which permits alternate glyphs to be selected and displayed (Microsoft Typography, 2008a). One drawback is that the person viewing the document must be using an application that supports these OT features in order for the specific alternate glyphs to appear. If the application does not support this OpenType feature, the default glyph for the Unicode character will appear. For example, the original author may have selected the shape “β” for U+03B2 GREEK SMALL LETTER BETA for his Greek text, but without the OpenType feature support, the recipient’s application may display a beta in the default shape, "ß".

OpenType also includes a way to specify specific glyph shapes that are commonly used for certain language-specific letters. For example, there are specific forms of italic and cursive Serbian letters that differ from Russian, although they are the same Unicode characters. The OpenType “language system” table and “locl” (localized form) feature table are mechanisms that allow one to specify such variant glyphs. These features are activated by language tags (Microsoft Typography, 2008b). However, as noted above, OpenType support – while becoming more common - is still limited to certain applications, although it is an international standard (ISO/IEC 14496-22:2009 [OFF] [ISO, 2011]).

4. Encode a Separate Character

One option occasionally mentioned as a way to represent a particular variant in a standardized way is to propose the variant as a separate character in Unicode. Technically, this is not allowed, since one of the core design principles is: “Unicode encodes characters, not glyphs” (Unicode Consortium, 2011a). However, some variants have been included in Unicode if they were present in earlier standards. The character/glyph model in CJK is particularly murky, in part due to the sheer number of characters involved (approximately 75,616 characters or 69% of all graphic characters in Unicode 6.0 are CJK). For the historic East Asian character sets, such as Classical Yi, the writing systems may be poorly understood and there is a tendency to encode glyphs. As a result, some character proposals have been based on glyphs (cf. the Classical Yi proposal, which proposed 88,613 “characters” [China, 2007]). Despite this, requesting

the encoding of glyph variants into Unicode (as separate characters) is not generally advisable.

5. Variation Sequence

A last option is to specify a Unicode variation sequence which is defined as a base character and a variation selector (Unicode Consortium, 2011b). This is a standardized means to indicate the glyphic variants of the base character. The advantage to this mechanism is that the variation is accessible in plain text, and does not rely on code points in the Private Use Area, which are not interoperable.

This particular mechanism has not yet been widely publicized amongst in the world of digital humanities. It will likely become more widely supported in software, particularly as the Japanese government will be using variation sequences to handle rare ideographs used in proper names and place names, rather than proposing 2,621 new “compatibility” characters (Japan , 2009). In 2010, the Japan National Body put forward a large collection of ideographic variation sequences, which have been under review (Unicode Consortium, 2010a).

Variation selectors have been mentioned as a way to handle variants for several historic scripts, namely Tangut and Manichaean. For Tangut, a historic script used in China until the 16c, the variation sequences were suggested as a way to handle cases where the lexical sources don’t agree (that is, there is disagreement whether a given glyph is a variant of a character or is a separate character), as a way to document when different scholarly opinions on unifications, and to address backwards compatibility issues (Cook and Lunde, 2008).

In Manichaean, the variation selectors are mentioned as a way to indicate alternate forms which are not predictable, either by their position in a word, or in a line. The use of the variation sequences maintains the basic character identity (Everson et al., 2009).

Figure 1 is an example showing the proposed shape for the Manichaean HE glyph, and the HE with Variation Selector-1. (See Figure 1)

One drawback is that the variation sequences need to be proposed and approved by the Unicode Consortium, much as new characters are (or, for ideographic sequences, are reviewed as part of the Unicode Public Review Process). However, this hurdle will ensure the characters are standardized, and are publicly accessible (Unicode Editorial Committee Members, 2011; Unicode Consortium, 2010b).

Another benefit is that search queries can ignore the variation selectors or the query can be written to only match a term with a specific variation selector. This mechanism could be useful as a way to display glyph errors, and be able to relate them to the base character. However, if a given application does not support variation sequences, the base character will display by default.

Variation sequences provide a standards-based option, which has some advantages over font-based alternatives. However, to date, relatively few variation sequences have been defined, except for those used in mathematics, Mongolian, and the historic script Phags-Pa (Unicode Editorial Committee Members, 2011).

At present, Ideographic Variation Sequences are only supported in the certain environments (Acrobat/ Reader 9.0 and higher, Flash Player 10 and higher, InDesign CS4 and higher, Mac OS X 10.6 and higher, Windows 7 and higher, and Firefox 4 on all platforms [Lunde, 2011]). The dependency on limited implementations can pose a problem for digital humanists, however, if future software fails to support these variation sequences.

6. Conclusion

In sum, several alternatives are available to text encoders to specify variant glyphs in text at present. This paper has provided new information on different options, which are still developing and may become more widely adopted, affecting choices available to text encoders.

This author cautiously recommends the use of Variation Selectors if the glyph difference needs to be captured in plain-text, and the digital encoder is willing to go through the approval process to get the variation sequence approved by the standards committees.

This work was supported by the National Endowment for the Humanities as part of the Universal Scripts Project [#PW-50441].

Figure 1: Manichaean HE glyph,
and the HE with Variation Selector-1

References

China [National Body]. (2007). *Preliminary Proposal to Encode Classical Yi Characters*. . . http://std.dkuug.dk/JTC1/SC2/WG2/docs/n3288.pdf (accessed 14 March 2011).

Cook Richard, Ken Lunde (2008). *The UCS Tangut Repertory.*. . http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3521.pdf (accessed 14 March 2011).

Everson, Michael, Desmond Durkin-Meisterernst, Roozbeh Pournader (2009). *Revised proposal for encoding the Manichaean script in the SMP of the UCS.*. . http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3644.pdf (accessed 14 March 2011).

International Organization for Standardization [ISO] (2009). . http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=52136&ics1=35 (accessed 14 March 2011)

Japan [National Body] (2009). *Follow-up on N3530 (Compatibility Ideographs for Government Use).* . http://std.dkuug.dk/JTC1/SC2/WG2/docs/n3706.doc (accessed 14 March 2011).

Lunde, Ken (2011). *E-mail to Deborah Anderson, 10 March.*

Microsoft Typography. (2008a). *Developer Info, OpenType specification, OpenType Layout tag registry: Registered features: Tag: 'cv01' – 'cv99'.*. http://www.microsoft.com/typography/otspec/features_ae.htm#cv01-cv99 (accessed 14 March 2011).

Microsoft Typography. (2008b). *Developer Info, OpenType specification, OpenType Layout tag registry: Registered features: Tag: 'locl'.*. http://www.microsoft.com/typography/otspec/features_ko.htm#locl (accessed 14 March 2011).

TEI Consortium, eds. (2010). *'5. Representation of Non-standard Characters and Glyphs.' In: TEI P5: Guidelines for Electronic Text Encoding and Interchange.*Version 1.9.1. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/WD.html (accessed 14 March 2011).

Unicode Consortium (2010a). *PRI 167: Combined registration of the Hanyo-Denshi collection and of sequences in that collection.*. http://www.unicode.org/ivd/pri/pri167/index.html (accessed 14 March 2011).

Unicode Consortium (2010b). *Ideographic Variation Database.*. http://www.unicode.org/ivd (accessed 14 March 2011).

Unicode Consortium (2011a). ''Chapter 2: General Structure.''. *Allen, Julie D., et al. The Unicode Standard Version 6.0 – Core Specification.*. Mountain View: Unicode Consortium. http://www.unicode.org/versions/Unicode6.0.0 (accessed 14 March 2011).

Unicode Consortium (2011b). ''Chapter 16: Special Areas and Format Characters.''. *In: Allen, Julie D., et al. The Unicode Standard Version 6.0 – Core Specification. Mountain View: Unicode Consortium.* http://www.unicode.org/versions/Unicode6.0.0 (accessed 14 March 2011).

Unicode Editorial Committee Members. (2011). *Standardized Variants. Revision 6.0.0.* . http://www.unicode.org/Public/UNIDATA/StandardizedVariants.html (accessed 14 March 2011).

W3C (2010). *Fonts on the Web.*. http://www.w3.org/Fonts/ (accessed 14 March 2011).

Supporting Scientific Discoveries to Answer Art Authorship Related Questions Across Diverse Disciplines and Geographically Distributed Resources

Bajcsy, Peter
pbajcsy@ncsa.uiuc.edu
National Center for Supercomputing Applications

Kooper, Rob
kooper@ncsa.uiuc.edu
National Center for Supercomputing Applications

Marini, Luigi
lmarini@ncsa.uiuc.edu
National Center for Supercomputing Applications

Shaw, Tenzing
twshaw3@ncsa.uiuc.edu
National Center for Supercomputing Applications

Hedeman, Anne D.
ahedeman@illinois.edu
University of Illinois

Markley, Robert
rmarkley@illinois.edu
University of Illinois

Simeone, Michael
mpsimeon@illinois.edu
University of Illinois

Hansen, Natalie
nhansen2@illinois.edu
University of Illinois

Appleford, Simon
sapplefo@illinois.edu
University of Illinois

Rehberger, Dean
dean.rehberger@matrix.msu.edu
MATRIX: The Center for Humane Arts, Letters, and Social Sciences Online, Michigan State University

Richardson, Justine
justine.richardson@matrix.msu.edu
MATRIX: The Center for Humane Arts, Letters, and Social Sciences Online, Michigan State University

Geimer, Matthew
matt.geimer@matrix.msu.edu
MATRIX: The Center for Humane Arts, Letters, and Social Sciences Online, Michigan State University

Cohen, Steve M.
steve.cohen@matrix.msu.edu
MATRIX: The Center for Humane Arts, Letters, and Social Sciences Online, Michigan State University

Ainsworth, Peter
p.f.ainsworth@sheffield.ac.uk
University of Sheffield

Meredith, Michael
M.Meredith@sheffield.ac.uk
University of Sheffield

Guiliano, Jennifer
jenguiliano@gmail.com
University of South Carolina

1. Overview

In the past, humanities scholars have primarily used text-based computational approaches to engage questions of authorship. In the area of visual arts, computational analysis of authorship is a growing field, but it is one that features diverse questions and requires complex algorithms, significant computational resources and a wide variety of experts from diverse disciplines to combine the results of visual inspections with computer generated results. Furthermore, when approaching the broad field of authorship-related questions in visual works, the variety of digital images representing cultural artifacts poses a formidable challenge on the robustness and accuracy of computer algorithms. The motivation of our work is to explore technologies that facilitate enquiry about authorship in visual art work and to address the challenges related to algorithm development, computational scalability of algorithms, distributed software development and data sharing, efficient communication tools across diverse disciplines, and robustness and general utility of algorithmic development when applied to a spectrum of authorship questions from historical images. In other words, we wanted to develop specific methods for new image-based research as well as build and model for future work in image processing and humanities research. We approached these challenges by selecting image subsets from the collections of 15th-century manuscripts, 17th and 18th-century maps, and 19th through 21st-century quilts

that often have corporate and anonymous authors working in community groups, guilds, artisan shops, and scriptoriums, and report technologies designed to support authorship discoveries in these collections. Crucially, the questions our algorithms and experts address are concerned with using authorship as a trope for analysis and generating new data, not just verifying the heritage or identity of a given artifact.

2. Methodology

The research being presented as part of this paper submission is derived from the Digging into Data to Answer Authorship Related Questions Grant awarded as part of the Digging into Data Challenge Competition (www.diggingintodata.org). An international, multi-disciplinary team of researchers from the University of Illinois (US), the National Center for Supercomputing Applications (US), Michigan State University (US), and the University of Sheffield (UK), the DID team works to formulate and address the problem of finding salient characteristics of artists from two-dimensional (2D) images of historical artifacts. Given a set of 2D images of historical artifacts with known authors, our project teams aim to discover what salient characteristics make an artist different from others, and then to enable statistical learning about individual and collective authorship. The objective of this effort is to learn what is unique about the style of each artist, and to provide the results at a much higher level of confidence than previously has been feasible by exploring a large search space in the semantic gap of image understanding. Team members are geographically distributed and have very different backgrounds and expertise. While the discoveries require involvements and interactions of experts in computer science and in humanities, we had to design a methodology for communicating, coordinating web design and public relationship interfaces, large size data sharing, collaborative software development, software sharing and testing, and hardware sharing. We approached this spectrum of collaborative project challenges by (a) establishing communication and coordination channels (ooVoo videoconference, mailing lists, legal point of contacts regarding licenses and intellectual properties), (b) designing and deploying a content repository called Medici, (c) designing and documenting a library of content based file comparisons with standard application programming interfaces (API) for software development called Versus, (d) deploying software source control and bug tracking systems accessible to all team members (SVN and JIRA), (e) designing web-based workflow systems that could give access to hardware resources at any site for execution of algorithms called Cyberintegrator, and (f) providing additional tools and user interfaces for humanity scholars to view large size images and contribute to the interpretation of the computer generated results.

3. Technical Approach and Initial Results

Emphasizing the aspects of data-sharing, collaborative software development, distributed hardware resources, and interactions of experts from diverse domains in the Digging into Data project, we designed, developed and deployed technologies supporting a wide spectrum of team activities. The data, software and hardware sharing technologies include the Medici Content Management Repository (see Fig. 1),

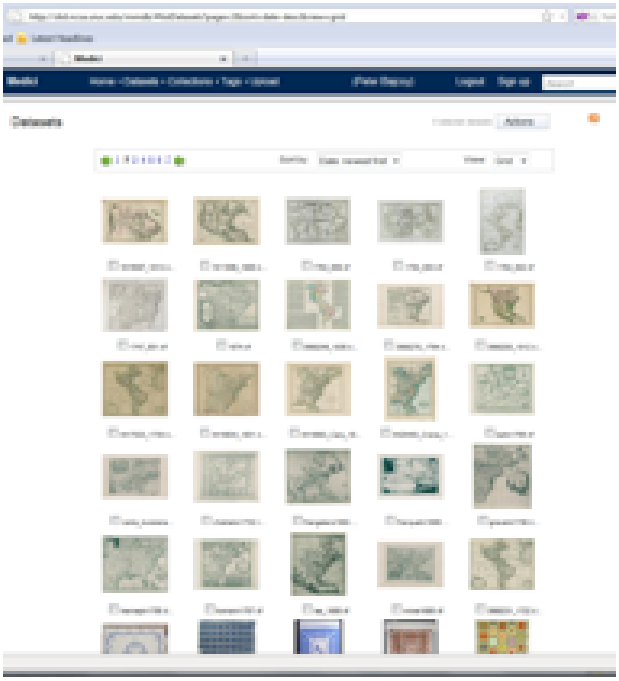


Figure 1: User interface to the Medici management repository for data sharing, annotations and visualization of large size images.

the Im2Learn library of basic image processing and visualization algorithms that can be applied to various image analyses (see Fig. 2),

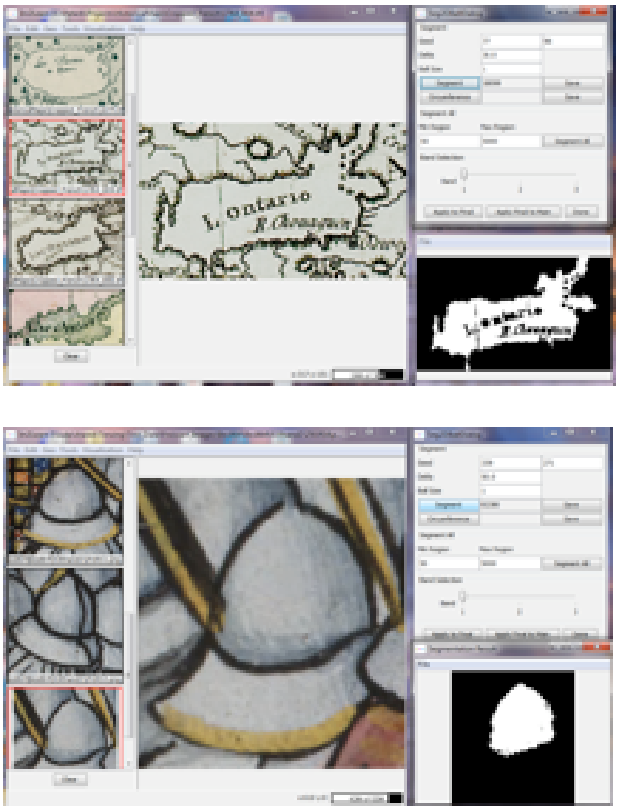


Figure 2: An example of a segmentation algorithm in Im2Learn library that was applied to historical map analyses (top) and manuscript illustration analyses (bottom).

the Versus library for content-based image comparison (see Fig.3),

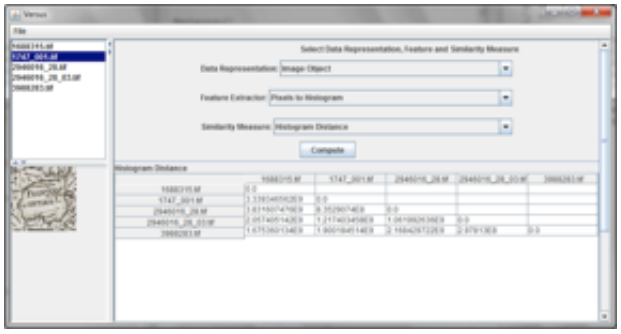


Figure 3: Initial user interface to Versus in order to support image comparison based analyses.

and the Cyberintegrator workflow for managing computations on distributed computational resources. The Medici Content Repository System is a web and desktop-enabled content management system that allows users to upload, collate, annotate, and run analytics on a variety of files types allowing for portable and open representation of data with extensible analytical tools. The analytical capabilities come from the Im2learn library that provides a plug-and-play interface for adding new algorithms and tools. Due to the fact that the authorship questions are frequently based on a comparison operation, we have designed additional API called Versus which allows everyone to contribute with comparison methods. Once the algorithms for image analyses and comparisons have been developed, they can be integrated into workflows (a sequence of algorithmic operations to reach the analytical goal) in Cyberintegrator workflow environment. Cyberintegrator is a user friendly editor to several middleware software components that:

1. enable users to easily include tools and data sets into a software/data unifying environment
2. annotate data, tools and workflows with metadata
3. visualize data and metadata
4. share data and tools using local and remote context repository
5. execute step-by-step workflows during scientific explorations
6. gather provenance information about tool executions and data creations.

In order to support visual explorations of large size images and contribute to the interpretation of the computer generated results by humanists and computer scientists, we have also integrated Microsoft Live Lab's Seadragon library to build image pyramids and support fast zoom in and out operations.

4. Summary

Our paper addressed each logistical and computational facet of a distributed, international collaboration. Based on our initial effort, all team members have responded positively to the technologies introduced and also helped in defining requirements for executing such complex projects involving collaborative humanities research. Based on our current observations, the web technologies for data, software and hardware sharing provided the foundation blocks for addressing the authorship discovery challenges. We have also concluded that the open nature of joint software development is necessary for overcoming intellectual property right and other legal hurdles.

5. Acknowledgment

We would like to acknowledge the NSF ITS 09-10562 EAGER grant and the NSF/NEH/JISC Digging into Data (NSF grant ID: 1039385).

Trailblazing through Forests of Resources in Linguistics

Barkey, Reinhild
rbarkey@sfs.uni-tuebingen.de
Department of Linguistics, University of Tübingen

Hinrichs, Erhard
erhard.hinrichs@uni-tuebingen.de
Department of Linguistics, University of Tübingen

Hoppermann, Christina
christina.hoppermann@uni-tuebingen.de
Department of Linguistics, University of Tübingen

Trippel, Thorsten
thorsten.trippel@uni-tuebingen.de
Department of Linguistics, University of Tübingen

Zinn, Claus
claus.zinn@uni-tuebingen.de
Department of Linguistics, University of Tübingen

1. Introduction

Linguistics is facing the challenge of many other sciences as it continues to grow into increasingly complex subfields, each with its own separate or overarching branches. While linguists are certainly aware of the overall structure of the research field, they cannot follow all developments other than those of their subfields. It is thus important to help specialists but also newcomers alike to bushwhack through evolved or unknown territory of linguistic data.

A considerable amount of research data in linguistics is described with metadata. While studies described and published in archived journals and conference proceedings receive a quite homogeneous set of metadata tags — e.g., *author*, *title*, *publisher* —, this does not hold for the empirical data and analyses that underlie such studies. Moreover, lexicons, grammars, experimental data, and other types of resources come in different forms; and to make things worse, their description in terms of metadata is also not uniform, if existing at all.

These problems are well-known and there are now a number of international initiatives — e.g., CLARIN, FlareNet, MetaNet, DARIAH — to build infrastructures for managing linguistic resources. The NaLiDa project, funded by the German Research Foundation, aims at facilitating the management and access to linguistic resources originating from German research institutions. In cooperation with the German SFB 833 research center, we are developing a combination of *faceted* and *full-text search* to give *integrated* access through heterogeneous metadata sets. Our approach is supported by a central registry for metadata field descriptors, and a component repository for structured groups of data categories as larger building blocks.

2. State of Affairs

An increasing number of research institutions in linguistics is systematically archiving research data and making such data publicly available. Users can access such archives via institution-specific websites or purpose-built software where resources can be searched and, in part, downloaded. Some institutions provide access to their archives via OAI-PMH so that the archives' public content can be harvested and fed into the metadata services of data centers in the community.

The metadata provided by the various institutions differ not only in quantity and quality but also in the format of description, as Fig. 1 illustrates. Typically, a research organization designs a metadata schema that it deems to serve best its institutional setting and the number and types of resources it hosts. Different organizations will likely yield various schemas with their own structure and terminology for the schemas' nodes. Such semantic heterogeneity may be complemented by syntactic heterogeneity as formats may vary (e.g., ASCII, relational database format, XML). An organization's resources can thus be seen as a forest of trees of the same kind, whereas a tree — i.e., a description of one resource — may have deformations at the leafs as the result of not adhering to the schema. Moreover, trees will look rather naked when resources are described sparsely.

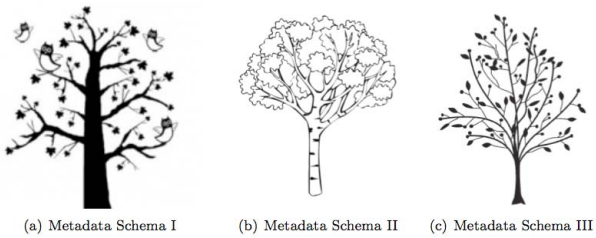


Figure 1: Metadata heterogeneity

3. Conceptual Setting

While we expect research organizations to continue managing their research data in their respective ways, we ask them (i) to make their data public

in a well-defined XML format to obtain syntactic uniformity, and (ii) to reformulate their schemas to adhere to CMDI (Component MetaData Infrastructure), see (Broeder et al. 2010), a component-based metadata model that makes use of predefined metadata components and the ISOcat data category registry (International Organization of Standardization 2009). The organizations' archive managers have the opportunity to redefine existing parts of their schema, but they can also choose to keep existing structures and terminologies. In this case the respective metadata descriptors need to be associated with their corresponding semantic points of references in ISOcat, being addressable via unique persistent identifiers. Moreover, archive managers can add new data categories to ISOcat's private space for immediate availability, and initiate a standardization process to pave the way for their wider use.

4. Technological Setting

4.1. Metadata Storage

The storage of resource descriptions has to cope with a multitude of schemas the descriptions adhere to. The use of a relational database would require a mapping of all schemas to a single one, which is all but trivial. Instead, with CouchDB [`http://couchdb.apache.org`], a no-SQL database is being used that stores arbitrarily structured documents rather than records of some fixed form. The translation between the XML-based CMDI-format into CouchDB's native JSON format is structure-and information-preserving.

4.2. Faceted Search

Facets serve to blaze the trail. Faceted search enables a user to find specific trees, i.e., resource descriptions, in the various forests by specifying (some of) their common properties. A facet partitions the search space where descriptions, i.e., CouchDB documents, in the same cluster share the same facet value. The selection of multiple facets corresponds to an intersection of clusters identifying resources that have all selected facet values. Faceted search also supplies a user with information about the number of documents in each cluster or intersections thereof — the "mileage" of following a trail. In the presently available data the following *unconditional* facets are adequate for the various schemas describing linguistic resources: "organization", "modality", "language", "country", "resourceType", and "origin". Facet values may stem from open or controlled vocabularies; controlled vocabulary facets have a stronger tendency to partition the search space into larger units, whereas open vocabularies induce larger search space fragmentations.

Once faceted search has focused on a subset of resources, *conditional* facets allow the introduction of additional context-specific navigational user aids. When users select the facet "resourceType" with value "tool", for instance, they restrict the search space to just encompass metadata that describes language-processing tools. Here, the conditional facet "toolType" is introduced that partitions the remaining search space according to the type of tool, e.g., language parser, spell checker. Moreover, conditional facets help lowering the complexity of computing search space clusters and their intersections.

4.3. Mapping Facets to Nodes of the Various Schemas

Metadata schemas vary in structure and terminology. Different names for nodes or leafs may be used to elicit the same meaning, and identical names may be used for semantically different concepts. This makes the mapping of facets to nodes of the various tree types (cf. Fig. 2) rather difficult, and usually requires some intricate knowledge of the metadata forests to be processed. When schemas make use of the aforementioned data category registry ISOcat, such ambiguities can be resolved automatically as names are linked to registry entries.

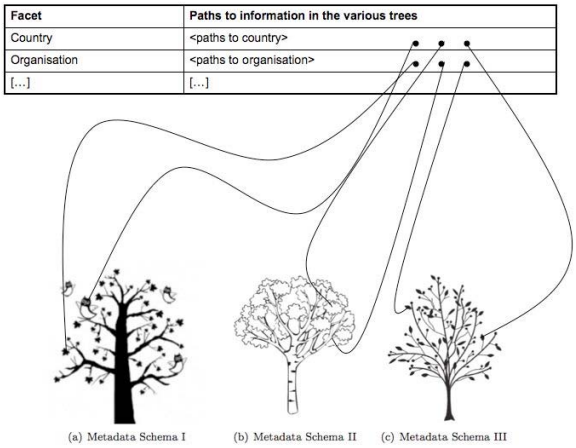


Figure 2: Mapping facets to the individual parts of the resource trees

Each facet corresponds to an elementary CouchDB view into the database of resource documents. These views serve as a starting point for the generation of complex views that correspond to the various possible navigational paths using facet selection, thus implementing the faceted browser back-end.

Elementary views and complex views are generated automatically from a facet specification file, an enriched textual encoding of the table in Fig. 2.

4.4. Full-text Search Support

Data sets may have resources that are hard to find using faceted search alone. This is true for resources with sparse metadata, or with descriptors that can rarely be mapped to facets. We are therefore using CouchDB's port to Lucene to perform full-text search across all resources or search spaces restricted by prior faceted search.

4.5. Front-end

Fig. 3 depicts a screenshot of the NaLiDa faceted browser; it shows a search state where users selected three facets ("country", "modality", "resourceType") and where the system displays an overview of the remaining search space in terms of the facets, the number of resources available for each of their values, and access to all documents selected so far.

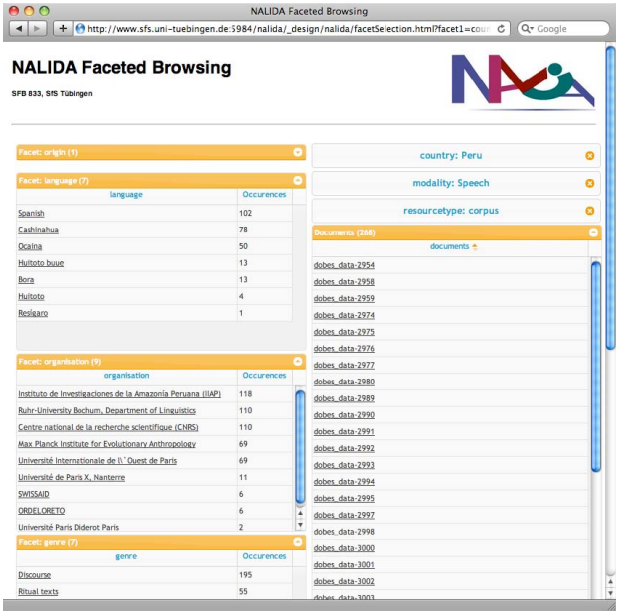


Figure 3: The NaLiDa Faceted Browser

5. Related Work and Conclusion

Faceted search is gaining popularity as users can explore large data sets without an intricate understanding of metadata fields or schemas; they obtain an immediate overview of the search space and guidance how to conquer it. A faceted search access to language resources has been implemened by the last author [`http://www.clarin.eu/vlo`] using Flamenco (Hearst 2006). Our new approach has four main advantages: CouchDB also stores the metadata documents (with varying schemas) and thus also serves as permanent storage; the use of conditional facets contributes to usability as only relevant facets are shown, guiding users' navigation; index generation accommodates for incremental updates on the metadata sets, supporting regular harvesting without recomputing all indices and views anew; and the faceted browser's back-end is generated automatically from a facet specification and can be configured easily for other datasets.

References

Broeder, D, Kemps-Snijders, M, Van Uytvanck, D, Windhouwer, M, Withers, P., Wittenburg, P, Zinn, C (2010). 'A Data Category Registry-and Component-based Metadata Framework'. *Proceedings of the 7th conference on International Language Resources and Evaluation, 19-21 May 2010.* European Language Resources Association.

International Organization of Standardization (2009). *Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources.* Geneva. `http://www.isocat.org`.

Hearst, M (2006). 'Design Recommendations for Hierarchical Faceted Search Interfaces'. *ACM SIGIR Workshop on Faceted Search*.

Lurking in Museums: In Support of Passive Participation

Smith Bautista, Susana
susanesm@usc.edu
University of Southern California

Lurking is a term that has gained popularity with the advent of online communities. Museums are no exception to this renewed interest in community building, both online and onsite. The Internet has helped museums to better serve their communities, connecting physical events and exhibitions with online services, information, and activities. The greater museum community is compartmentalized into different groups, including educators, scholars, teenagers, families and more, but their dues-paying members are perhaps the closest to what is more commonly known as *affinity spaces* (James Gee) or knowledge-sharing communities of practice (CoP). Many of the membership groups organize events at the museum, raise funds for the museum, socialize regularly, and even have online profiles, blogs, or pages on social media sites. The term *lurking* has arisen because online communities have high expectations for their members to participate and contribute, in particular with online games and chat forums. The web usability expert Jakob Nielsen (2006) proposed the well-known 90-9-1 rule of user participation in online communities, which states that 90% of users are lurkers, 9% of users contribute from time to time, and 1% of users account for most contributions. The more passive acts of being present (virtually or physically), listening, watching, and reading – that is, lurking – are considered negative when contrasted to the more dynamic acts of writing, contributing information, performing tasks, or discourse that are all viewed as essential to the formation and maintenance of community. Even the spectatorial can be considered negatively, as in *voyeurism* that is often perceived as leading to perverse and criminal acts.

Analogous to the conventional concept of community is the idea of a social network as a system of individual nodes that are all related, first proposed by J. A. Barnes in the early 1950s. The concept of *the network society* has been best developed by Manuel Castells (2000, p. 12) who states that, “The ability of an actor in the network – be it a company, individual, government, or other organization – to participate in the network is determined by the degree to which the node can contribute to the goals of the network… This leads to a binary process of inclusion and exclusion from the network. The people at the bottom are those who, with nothing to offer the network, are excluded.” Similar to lurking but more related to economics is the *free rider* problem, as it can be argued that every community – regardless of its size or nature – offers a public good and is based upon some type of exchange system.

Art museums are spaces that have traditionally encouraged lurking, as visitors are invited to leisurely appreciate works of art in a reverent environment that prioritizes observation, contemplation, learning, and personal interpretation. There are some exceptions, however, such as with participatory art practices that gained prominence in the early 1960s and continue to be exhibited in museums today, dependent on visitors’ active participation for their realization. Museums have entered the digital age just as have other traditional socio-cultural institutions, and consequently they are incorporating new technologies for the purposes of facilitating exhibition, interpretation, education, and participation. The modern museum presumes that visitors – especially younger ones (*digital natives*) – expect a more interactive museum experience that allows them to actively participate and even share their opinions within a community that is becoming perceptibly less hierarchical and authoritarian. New technologies offer tremendous possibilities for all visitors to engage more deeply with art, but they can also distract from the passive acts of contemplation and observation if they demand too much physical interaction. Nevertheless, art museums remain a trusted and respected place in which to observe, think, feel, and learn. This paper will assert that lurking is a necessary and useful part of community engagement and learning in the digital age, and that art museums are a valuable and unique space for such activity. Some specific technologies used by art museums today will be discussed in the context of whether they promote a more active or passive experience. A critical distinction must also be made, however, between the terms interaction and participation, the latter of which is more open-ended than the former.

In writing about participatory culture and the digital age, many scholars such as Henry Jenkins and Mikuzo Ito discuss the importance of participation amongst youth while also stressing the importance of other skills, characteristics, and stages of learning that could easily be construed as lurking. We understand from Richard Bartle’s (1996) taxonomy, D.T. Schaller et al.’s (2007) four learning preferences, and Ito et al.’s (2008) three genres, that there are many ways to engage

in activities, including both play and learning. The more active forms include creating, producing, sharing, contributing, playing, and commenting. These forms of participation are the most visible to the community and as such, are most prized in that they serve as an example for other members to emulate. More passive forms include listening, reading, watching, and browsing, as well as the introspective acts of thinking, reflecting, evaluating, and forming opinions. When these passive acts are shared with other individuals, discourse arises (physical or virtual), which can then be considered a more active and public form of participation. Different personalities also emerge within groups, and more active contributors will try to lead others less inclined to participate. Jenkins et al. (2008, p. 7) assert that, "In such a world [participatory culture], many will only dabble, some will dig deeper, and still others will master the skills that are most valued within the community." This paper will show that all types of participation are essential to the development and maintenance of communities. The formation of knowledge and learning is a linear process that begins with more introspective and often individual acts, which in many instances then become public when thoughts, creations, and knowledge are shared with others. If thoughts are not shared and knowledge remains private, then the learning process merely continues with the individual.

The negative implications of lurking arise mostly from within the communities themselves. What cultural communities tend to forget, however, is that lurkers play an important role; they provide an audience, they carefully observe community norms and practices, and they contemplate and interpret that which they observe to perhaps share later with others. Performers do not want to perform without an audience, writers do not want to write books that nobody reads, and museums cannot open without visitors. Certainly, artists have a personal drive to create that does not require an audience or even a client, but even the most dedicated artist desires feedback, validation, or public acknowledgment. Physical spaces and organizations have long measured success in quantitative terms of the number of seats filled, visitors through the front door, or books or tickets sold. Online spaces are no different; the success of websites are most commonly measured by the number of clicks, page views, or downloads, rather than the more interactive number of comments, links, or uploads. As Web 2.0 features become more common on websites, these interactive metrics will surely begin to matter more. Still the number of people "passively" watching, reading, and listening are a strong measure of success, as well as a strong incentive for creators, organizations, and funders alike.

Museums need to protect the act of lurking. Any expectation of active participation, interaction, and sharing may inhibit lurkers from eventually participating, for lurkers need not be defined as having rigid characteristics but rather as representing merely one stage in the long process of learning and civic engagement. In its recent report *Museums & Society 2034: Trends and Potential Futures*, the American Association of Museums (2008, p. 19) states that, "While technological progress has brought much value to society, one byproduct of these emergent structural shifts in communication technologies is almost certainly going to be a world with fewer and fewer places where the public can find respite and retreat." Through their expertly researched and curated exhibitions and related public programming, museums are best able to teach their visitors and members how to observe, how to critically think, and how to develop opinions in order to more effectively act on them if so desired. They can also teach the value of being an audience within a socially networked environment, which is the first step to recognizing the importance of community and public goods. This paper will discuss how museums might continue to encourage lurking in synergy with new digital technologies, and likewise how youth can become empowered to not only act, but also to observe and contemplate. As Jenkins et al. surmise (2008, p. 39), "…knowing how to act within the distributed knowledge system is more important than learning content. Because content is something that can be 'held' by technologies such as databases, websites, wikis, and so forth, the curricular focus is on learning how to generate, evaluate, interpret, and deploy data." Those who lurk in museums (online or physically) are not evading their role; they do have a very important role within their community and within the process of acquiring, forming and sharing knowledge, and even more so in the participatory culture of the digital age.

ComPair: Compare and Visualise the Usage of Language

Beavan, David
David.Beavan@glasgow.ac.uk
University of Glasgow, United Kingdom

1. Introduction

This paper will demonstrate ComPair, a new tool to investigate and compare word usage, encouraging new ways to explore language variation. While remaining focussed on the usability and the promotion of navigation, this tool represents an evolutionary step forward from the author's previous award winning visualisation applications. This paper will introduce the methods and technologies at its core, perform a demonstration of the tool and discuss opportunities for further collaboration.

2. Collocation

Firth in 1957 tells us 'You shall know a word by the company it keeps' leading to a contextual investigation of language which remains with us today. Identifying a word of interest and examining its collocates, often tells us more than a traditional dictionary definition ever could. Traditional corpus tools display collocates in tabular format, providing rich statistical data at the expense of giving the user an opportunity to see the overall linguistic landscape. Tools such as Beavan's Collocate Clouds present this information very differently, visualising the collocates in cloud form, as in figure 1.

Figure 1. Collocate Cloud of node word 'stars' [http://www.scottishcorpus.ac.uk/corpus/bnc/collocatecloud.php?word=stars (accessed 1 November 2010)]

A collocate, if known, can be quickly located due to the alphabetical nature of the display. Frequently occurring collocates stand out, as they are shown in a larger typeface, with collocationally strong pairings highlighted using brighter formatting. Therefore bright, large collocates are likely to be of interest, whereas dark, small collocates perhaps less so.

3. Comparison

Louw introduced us to semantic prosody, which describes how synonymous words can actually take on positive or negative connotations. A natural way to investigate this would be to separately compare the collocates of each node word of interest. This can be performed by looking at multiple collocate clouds side by side, or by using statistical tools presenting tabular data. While these methods may be best suited to the comparison of many node words of interest, ComPair provides a solution to the comparison of two words, while keeping true to the aims of collocate clouds.

Semantic prosody is illustrated in figure 2, comparing the collocates of 'utterly' vs. 'absolutely'. Negative terms cluster near 'utterly' where as positive terms cling to 'absolutely'. At face value these words are synonymous, but they are clearly used in different contexts and are not simply interchangeable. These are often issues which challenge learners of English as a foreign language.

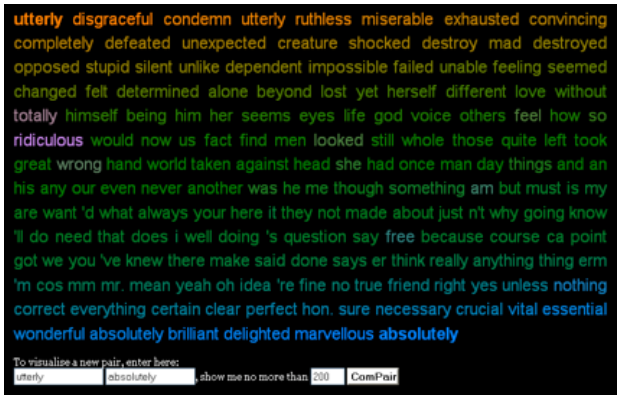


Figure 2. ComPair visualisation of 'utterly' vs. 'absolutely' in the British National Corpus

4. Method

ComPair calculates the collocates of both node words, ranking the results using a combination of frequency of co-occurrence, and by collocational strength (adopting the Mutual Information (MI) measure). A continuum is formed, with each extremity representing the separate node words (imagine a piece of string, with a label representing each search term at each end). The collocates are then distributed along this continuum, using the relative pull of collocational strength towards each node (the words near the end of the string are

strongly associated with those end labels, those in the centre much less so).

5. Visualisation

ComPair displays this continuum, displaying each node word and the collocates between them. The display uses a spectrum of colour, to further enforce the ordering of the collocates. In the MI tug of war, the words in the centre share similar MI scores with each input term. Typically these are fairly low MI figures and appear green. In the ‘utterly’ vs. ‘absolutely’ example above ‘ridiculous’ appears in pink, this indicates that while the MI scores (utterly- ridiculous’ vs. absolutely-ridiculous) are roughly the same, they are much higher than the surrounding collocates. Ridiculous is therefore a word used strongly with both utterly and absolutely. Those collocates appearing close to each node, and sharing its colour are used very strongly with that node word, and only that word. Figure 2 tells us that things can be ‘absolutely marvellous’ but not ‘utterly marvellous’. In comparison, someone can be ‘utterly ruthless’, but not ‘absolutely ruthless’.

6. Future Directions

At present ComPair allows for the comparison of two separate words in a single corpus. One possible extension would be the facility to search for the same words across two corpora. Imagine two corpora of differing political parties. With a single search term, ComPair would help expose the views and attitudes towards that concept. Another avenue would contrast word usage in British vs. American English.

Other applications would involve its use as a learning tool, allowing users to go beyond dictionaries and thesauri, to see in detail how different words actually operate. Visualisation of more than two node words should also be possible given different display techniques.

References

Beavan, D. (2008). 'Glimpses though the clouds: collocates in a new light'. *Proceedings of Digital Humanities 2008.* University of Oulu, 25-29 June 2008.

Firth, John R. (1957). *Modes of meaning.* Oxford: Oxford University Press.

Louw, B. (1993). 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies'. *Text and Technology [ed. Baker, M., Francis, G. & Tognini-Bonelli, E.].* Philadelphia/ Amsterdam: John Benjamins.

gMan: Creating General-Purpose Virtual Environments for (Digital) Archival Research

Blanke, Tobias
tobias.blanke@kcl.ac.uk
Centre for e-Research, King's College London

Connor, Richard
richard.connor@cis.strath.ac.uk
University of Strathclyde

Hedges, Mark
mark.hedges@kcl.ac.uk
Centre for e-Research, King's College London

Kristel, Conny
c.kristel@niod.knaw.nl
Netherlands Institute for War Documentation (NIOD)

Priddy, Mike
priddy@mac.com
Centre for e-Research, King's College London

Simenoni, Fabio
fabio.simeoni@cis.strath.ac.uk
University of Strathclyde

This paper will present a critical analysis of our attempts to build Virtual Research Environments (VREs) for everyday Humanities research tasks using digital archives. Numerous specialised VREs have been developed for addressing particular tasks in various humanities disciplines. The Silchester VREs addressed data integration in archaeological excavations, the SDM VRE developed services for sharing and annotating manuscripts, while TEXTvre is concerned with TEI-based resource creation. Building on these experiences, gMan addressed the issue of moving beyond support for specific, focused tasks, and instead building services to enable more general-purpose humanities research activities, such as integrating and organising the heterogeneous and often unstructured digital resources, and support for 'active reading' processes[1] through advanced discovery facilities. Such services regularly top the list of humanities user requirements[2]. This paper describes work to this end, firstly by the DARIAH project, and subsequently consolidated by the gMan project, funded by JISC's VRE Rapid Innovation programme.

These experiments were based on use cases identified by the earlier LaQuAT (Linking and Querying Ancient Texts) project[3], which investigated how to integrate scattered, heterogeneous and autonomous data resources relating to ancient texts, mainly databases but also XML corpora. LaQuAT attempted to solve these issues by offering an integration framework based on the OGSA-DAI grid middleware, which provided an integrated interface to the various data resources that followed a relational database model. However, this approach had certain limitations for our purposes, as such models are optimised for dealing with datacentric resources - that is, resources consisting primarily of structured data such as numbers, dates or very short text fields - rather than text-centric resources containing significant quantities of unstructured text. The approach worked well where the structural context of the information was clear and the query aimed at exact matches. More commonly, however, humanities researchers work with text-centric resources, perhaps enhanced with XML mark-up to capture document structure and additional metadata[4], and they look for resources for further investigation based on looser criteria of relevance, e.g. by searching for all Roman legal texts in one resource containing information on punishments that are also mentioned in papyri from another resource.

These conclusions were further elaborated in the use cases that were developed from them, which are the main drivers for the work described here. Complementing this is a body of methodological investigation concerning scholars and their use of sources, particularly their use of data and archives. Before describing our current work, we will survey briefly these investigations.

The difference in scholarly practices between the sciences and the mainstream humanities is highlighted in a study[5] that investigated the types of information sources used in different humanities disciplines, based on results from the US Research Libraries Group reports. Structured data is relatively little used, except in some areas of historical research, and data as it is traditionally understood in the sciences, e.g. the results of measurements, even less so. It is true that the study is partly outdated, and that data in the traditional sense is increasingly important in the humanities, particularly in linguistics and archaeology where scientific techniques have been widely adopted. Nevertheless, it is clear that in general humanities research relies not on measurements as a source of authority, but rather on the provenance of sources and peer-assessment, and that what data repositories are for the sciences, archives are for the humanities[6].

Archival records are primary sources about the past and may take many forms, including government papers, financial documents, photographs, sound recordings, etc. All this information is unstructured in nature.

Thus, our work is driven partly by the requirements from[7], interpreted so as to relate to methods of research in archives. Retrieval is to happen in real time, and traditional finding aids are to be complemented by more sophisticated retrieval mechanisms, including the ability to create relevance indexes on unstructured resources, as well as the ability to combine resources in new ways. In particular, we aimed to implement the personal copy of a finding aid that is often quoted as an important prerequisite for specialised research in archives.

Our work investigated how (digital) archival content can be delivered to humanities researchers more effectively, independently of the location and implementation of that content, and with special facilities provided for customising the retrieval, management and manipulation of the content. We investigated how the UK and European research infrastructure (RI) can be exploited to support data-driven, collaborative research in the humanities by using the gCube environment[8], which was developed by the EU-funded D4Science project. gCube allows virtual research communities to deploy VREs on demand by making use of the shared resources of the European RI, and provides services that match closely the sort of information organisation and retrieval activities that we identified as being typical in humanities research.

D4Science provides an easy way of scavenging online data resources. It has a consistent mechanism to import data for rich user interaction within the deployed VREs. Its data resource staging framework, based on a well-defined workflow of data analysis, data modelling and data generation, is one of the key innovations of D4Science. The analysis and modelling phases define how data collections are loaded into gCube compound objects using its simple but powerful data model. In the data generation phase, descriptive metadata and provenance information are added.

Using the gCube data staging framework, the following datasets were brought together in our experiments:

- The Heidelberger Gesamtverzeichnis (HGV) der griechischen Papyrusurkunden Aegyptens, a collection of metadata records for 65,000 Greek papyri from Egypt.
- Projet Volterra, a database of Roman legal texts, currently in the low tens of thousands but very much in progress, stored in a series of themed tables in Microsoft Access.
- The Inscriptions of Aphrodisias, a corpus of about 2,000 ancient Greek

These datasets were the same as those used in the LaQuAT project and thus and allow a critical comparison of results. They overlap in terms of time, places and people – specifically looking at the first five centuries or so of the Roman Empire – although their contents are otherwise quite different. The provision of an environment for working with this data in an integrated form would be highly fruitful for the researcher.

The presentation will describe the use cases that we used for evaluating gCube. Our approach was to break down the scenarios identified in interviews at KCL and within DARIAH into a number of common, atomic actions. Specific instances of these actions can be combined to model a variety of "real" research scenarios, for example the ability to assemble heterogeneous resources (or parts of resources) into a virtual collection, to share this virtual collection within a specific community and to search across a virtual collection, where specific search parameters (such as the importance of specific locations) can be set according to preference. Specific communities also require specific search services such as geo-referenced and date-range searches. Finally, the researcher wants to share links between research objects and annotations (including related documents publications) in her community.

In our experiments, we confirmed that most of these use cases could be supported by the features already provided by the core D4Science systems. For the Digital Humanities 2011 presentation, we will address our subsequent activities: the analysis that we carried out to identify gaps in the existing service provision; some results that demonstrate a clear distinction between the viewpoints of humanities and science research, in respect of such features as image search; our move to develop gMan as a production service for humanities researchers; and the recently-funded European Holocaust Research Infrastructure project, which aims to integrate Holocaust research material from archives across Europe. The main aim of the project will be to make accessible existing Holocaust research collections but the second priority will be to deploy virtual research environments to make use of these resources. D4Science services were seen to support initial requirements well.

Notes

1. Brockman et al. 2001. Scholarly work in the humanities and the evolving information environment. Washington, DC.
2. Benardou et al. 2009. Understanding the Information Requirements of Arts and Humanities Scholarship. International Journal of Digital Curation.
3. Jackson et al. 2009. Building bridges between islands of data—an investigation into distributed data management in the humanities. Proceedings of the Fifth IEEE International Conference on e-Science. Washington, DC.
4. Nentwich, M. 2003. Cyberscience. research in the age of the internet. Vienna.
5. Palmer et al. 2009. Scholarly information practices in the online environment.
6. Duff et al. 2004 Historians' use of archival sources: Promises and pitfalls of the digital age. The Public Historian.
7. Duff et al. 2004 Historians' use of archival sources: Promises and pitfalls of the digital age. The Public Historian.
8. Candela et al. 2009. On-demand Virtual Research Environments and the Changing Roles of Librarians. Library Hi Tech.

Topic Modeling Historical Sources: Analyzing the Diary of Martha Ballard

Blevins, Cameron
cblevins@stanford.edu
Stanford University

In 1991, historian Laurel Ulrich's *A Midwife's Tale* swept a little-known 18th-century midwife named Martha Ballard into the national historical consciousness. Ulrich's work centered on the analysis of nearly 10,000 diary entries penned by Ballard between 1785 and 1812, leading to an exploration of issues such as shifting family structures, the professionalization of obstetrics, and debtor patterns in a rural economy (Ulrich 1991). My research examines the same diary, but instead of a traditional close reading of the source, I use topic modeling to mine a digitized transcription, iterating through hundreds of thousands of words in order to search for textual patterns.

One of the fundamental challenges to applying text processing techniques to historical sources is one of data quality. Older, hand-written documents are often difficult to transcribe into a digital format, while the shorthand style of diary writing is often filled with abbreviations and misspellings. For instance, Ballard employs a vocabulary peppered with variations: the word "daughter" is spelled fourteen different ways: "daught," "dagt," "dat," etc. One way to overcome this challenge is to use topic modeling, a method of computational linguistics that attempts to group words together based on their appearance in the text.

My short paper session focuses on an analysis of a historical source using topic modeling (Blei and Lafferty 2009). As a form of linguistic analysis, topic modeling has been employed over the past several years to examine large-scale, multi-author textual databases, including historical newspapers (Block 2006), journal articles (Gerrish et al. 2010, Hall et. al. 2008), and social network data (Ramage et al. 2010). My application of topic modeling differs from many of these investigations by focusing on multiple, short texts by a single author: in this case, Ballard's diary entries.

I employed the machine learning toolkit MALLET (McCallum 2002) in order to topic model each of Ballard's entries as separate pieces of text. MALLET, identified thirty topics, which I then labeled for clarity.

The following sample topics were some of the most coherent (my own labels in bold and uppercase):

Topic Label	Topic Words
MIDWIFERY	birth deld safe morn receivd calld left cleverly pm labour fine reward arivd infant expected recd shee born patient
CHURCH	meeting attended afternoon reverend worship foren mr famely performd vers attend public supper st service lecture discoarst administred supt
DEATH	day yesterday informd morn years death ye hear expired expird weak dead las past heard days drowned departed evinn
GARDENING	gardin sett worked clear beens corn warm planted matters cucumbers gatherd potatoes plants ou sowd door squash wed seeds
SHOPPING	lb made brot bot tea butter sugar carried oz chees pork candles wheat store pr beef spirit churnd flower
ILLNESS	unwell mr sick gave dr rainy easier care head neighbor feet relief made throat poorly takeing medisin ts stomach

Topics

Although topic modeling was useful for overcoming some of the challenges of spelling variations, its real value lies in its ability to quantitatively measure the relative thematic content of each piece of text. In the case of Ballard's diary, MALLET assumes that each diary entry is compromised of some combination of thirty topics. An entry in which Ballard attended a sermon and purchased supplies from the general store might contain, for instance, scores of 50% for the CHURCH topic, 25% for the SHOPPING topic, and minimal or zero scores for the remaining twenty-eight topics. Associated temporal metadata (day, month, year, day of the week) allowed me to chart the behavior of certain topics over time.

As a simple barometer of its effectiveness, I used one of the generated topics that I labeled COLD WEATHER, which included words such as cold, windy, chilly, snowy, and air. Aggregating its entry scores by month shows exactly what one would expect over the course of a year (Figure 1).

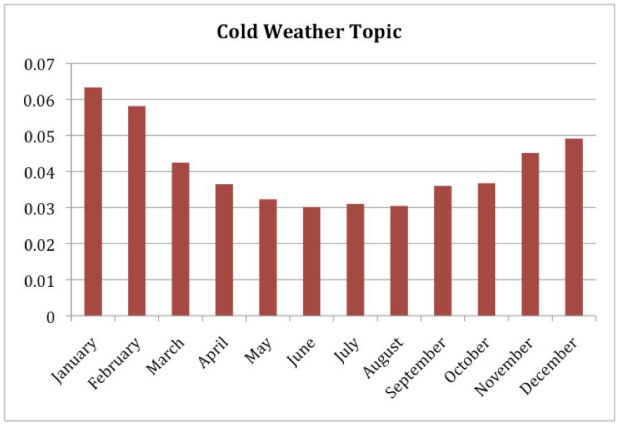


Figure 1

This approach also can chart patterns over the course of the diary, which covers the final twenty-seven years of Ballard's life. Two topics tended to involve words related to HOUSEWORK. Aggregated by year, they demonstrate a steady increase in the frequency with which Ballard writes about daily chores (Figure 2).

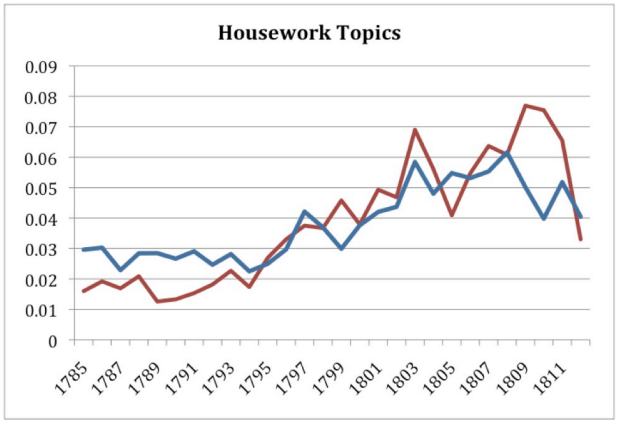


Figure 2

Both topics moved in tandem and steadily increased as she grew older (excepting a curious divergence in the last several years of the diary). This is somewhat counter-intuitive, as one would assume the household responsibilities for an aging grandmother with a large family would decrease over time. Yet this pattern bolsters the argument made by Ulrich in *A Midwife's Tale*, in which she points out that the first half of the diary was "written when her family's productive power was at its height." (Ulrich 1991, pp. 285) As her children married and moved into different households, and her own husband experienced mounting legal and financial troubles, her daily burdens around the house increased. Topic modeling quantifies and visualizes this pattern, one not immediately visible to a human reader.

Topic modeling allows for patterns to crystallize that are imperceptible to a human reader. One topic was particularly intriguing, and included the words: feel husband unwel warm feeble felt god great fatagud fatagued thro life time year dear rose famely bu good

These were words that seem to cover EMOTION and spiritual reflection – an abstract topic that is difficult enough for a human reader to describe. Yet the computer did a remarkable job in identifying a cohesive group of words. The topic follows a fascinating trajectory in Ballard's diary (Figure 3).

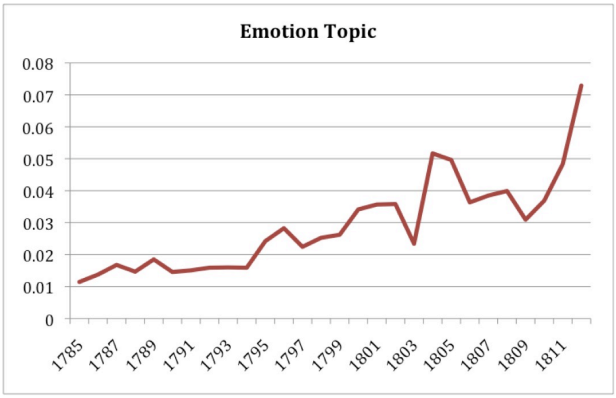


Figure 3

Not only did Ballard write about this topic more as she grew older, but there was a dramatic leap from 1803 to 1804-1805. This corresponds quite well to the period of intense family travail: Her husband was imprisoned for debt and her son was indicted by a grand jury for fraud, causing a cascade effect in Martha's own life. Topic modeling not only reveals the trajectory of tangible themes (housework, births, gardening, etc.), but also begins to quantify and visualize abstract themes by charting Ballard's emotional state of being.

My short paper session focuses on the results of my existing work on topic modeling Ballard's diary while outlining some of the future paths this research could take. In particular, I am interested in pairing trends in topics with trends in Ballard's social network. What topics correlate with what kinds of people? Are women or men described alongside particular themes? In what broad context do ministers, doctors, neighbors, or family members appear? In conjunction with traditional research and analysis, topic modeling presents a valuable methodology for examining historical sources.

References

Blei. D., Lafferty, J. (2009). 'Topic Models'. *Text Mining: Classification, Clustering, and Applications.*. Srivastava, A. and Sahami, M. (ed.). Boca Raton: Champan & Hall, pp. pp. 71-94.

Block, Sharon (2006). 'Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources'. *Common-Place.* 6.2. http://www.common-place.org/vol-06/no-02/tales/.

Gerrish, S., Llewellyn, C. (2010). 'JSTOR Discipline Browser'. *JSTOR.* http://showcase.jstor.org/projects/discipline-browser.

Hall, D., Jurafsky, D., Manning, C. (2008). 'Studying the History of Ideas Using Topic Models'. *Proceedings of Empirical Methods of Natural Language Processing.*. http://nlp.stanford.edu/pubs/hall-emnlp08.pdf.

McCallum, Andrew (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Ramage, D., Dumais, S., Liebling, D. (2010). '"Characterizing Microblogs with Topic Models."'. *International Conference on Weblogs and Social Media.*. http://nlp.stanford.edu/pubs/twitter-icwsm10.pdf.

Ulrich, Laurel Thatcher (1991). *A Midwife's Tale: The Life of Martha Ballard and Her Diary, 1785-1812.*. New York: Vintage.

Cinemetrics: A Digital Laboratory for Film Studies

Bosse, Arno
abosse@uchicago.edu
The University of Chicago

Tsivian, Yuri
ytsivian@uchicago.edu
The University of Chicago

Brisson, Keith
brisson@uchicago.edu
The University of Chicago

Cinemetrics (www.cinemetrics.lv) is a collaborative, online tool to enable researchers to collect, store, and process scholarly data about film editing. The long-term goal of the project is to create an extensive, multi-faceted collection of freely accessible digital data on film editing - a digital laboratory for the study of film style. Over the last four years, over 600 individuals have contributed editing data on circa 6,900 films to the project. Our paper at DH2011 will focus on the types of film historical questions Cinemetrics can address today. In addition, we'll provide a brief overview of how the software and the collaborative submission works, compare our work with similar efforts, and finally offer an early look at future directions for development.

At present, Cinemetrics is programmed handle the aspect of editing known in film studies as cutting rates. Though we tend to perceive their unfolding as continuous most films consist of segments called shots that are separated by instant breaks called cuts. Shots differ in terms of space and in terms of time. We know enough about space-related distinctions between shots, which are easy to name (shot 1: baby playing; shot 2: man looking) and categorize (shot 1: medium long high angle shot; shot 2: facial close up). Time-related differences between shots are more elusive and harder to talk about for; unlike in music or poetry with their scaled feet and measures, variations in shot length are not ones of distinction but of degree.

Shot lengths are sometimes convenient to present as the frequency of shot changes, or cuts, hence the term "cutting rates." Shorter shots mean a higher cutting rate. Unsurprisingly, cutting rates are linked to the story and its space-time articulations; likewise, montage sequences meant to cover larger spaces of story time have higher cutting rates than sequences shown in real time. Less evident, but just as important, is the relationship between cutting rates and the history of film. This is this gap in our knowledge that Cinemetrics is designed to bridge.

Using Cinemetrics, we are able to obtain and present cutting-related data in a more flexible way than previously available. Rather than calculate average shot lengths (ASL) arithmetically, Cinemetrics records and stores the time-span of each separate shot. Distinct from the arithmetical ASL, which is a single datum, Cinemetrics treats each film as a database of shots and highlights its individual features. Specifically, it tells us about a film's cutting swing (standard deviations of shorter and longer shots from ASL), its cutting range (difference in seconds between the shortest and the longest shot of the film), and its dynamic profiles (polynomial trendlines that reflect fluctuations of shot lengths within the duration of the film). In the "Articles" section of our site are links to articles by a number of film scholars on movie measurement studies using Cinemetrics.

The Cinemetrics database is an open-submission repository of data collected by people who use the client tool. All raw research data submitted to the site is freely available to anyone. The database's default sorting is alphabetic by film titles, but it can also be sorted by other parameters such as year, submitter's name, submission date, simple vs. advanced mode of measuring, and by the film's average shot length, median shot length, and standard deviation. By clicking on a film, title the user can access the page that provides basic statistics and interactive graphs related to this film.

"Cinemetrics Lab" is the latest addition to our site, and a work in progress. It is envisaged to offer the students of film history a range of analytical tools that will help them dissect, visualize, and compare film-related data. We started with a large-scale comparative map that looks a little like a star map. It is a scatter graph, and each dot represents a film available on our database. If you find your film on this map, you will instantly see how it relates to thousands of other films on the x-axis on time (111 years of film history) and on the y-axis of average shot lengths.

While Cinemetrics has no clones, there are a number of projects pursuing similar goals that complement our efforts. Jeremy Butler's useful "Shot Logger" (www.tcf.ua.edu/slgallery/shotlogger/) features a database of films (mainly TV) and offers statistics "inspired by Cinemetrics", but its database is still small, and the seven statistics values the site yields are numerically, not

graphically, expressed. The francophone site "Lignes de Temps" (web.iri.centrepompidou.fr/pop_site.html), linked to the Georges Pompidou Center for Modern Art in Paris is mainly designed as a video-flow annotation and cut-detection tool.

The above-mentioned "Shotlogger," "Edit 2000," and especially "Research into Film" are three sites on which Cinemetrics activities are actively echoed or discussed. Nick Redfern's "Research into Film" (nickredfern.wordpress.com/) uses Cinemetrics data to theorize statistical approaches to film studies. "Edit 2000" (www.data2000.no/EDIT2000/), launched in Norway in 2009, was made to represent Edit Decision List (EDL) files as numeric and visual summaries. Another group whose researchers deployed Cinemetrics raw data for their experiments in data visualization is "Software Studies Initiative" headed by Prof. Lev Manovich at UC San Diego. The site shows how Cinemetrics data can be variously represented using different visualization tools.

The Digital Archaeological Record--an Analytic Data Repository for Archaeology

Brin, Adam
abrin@digitalantiquity.org
Digital Antiquity

McManamon, Francis
fpm@digitalantiquity.org
Digital Antiquity

Lee, Allen
Allen.Lee@asu.edu
Arizona State University

In the past 150 years, the discipline of archaeology has changed dramatically; excavation procedures, field methods, and record keeping have both improved and become formalized. Ahead of the digital era, the physical records of an excavation (the papers, data tables, journals, and monographs) were preserved as artifacts alongside the excavated materials in museums and repositories. More recently, Archaeologists have been quick to adopt new technology from punch cards in the 1960s to spreadsheets, databases, GIS, and 3D scanning. Yet, these modern files, images, data sets, and documents, if not properly preserved, are more fragile than the objects they describe. With the Digital Archaeological Record (tDAR --http://www.tdar.org), we hope to change this.

tDAR was designed as a domain-specific digital repository, focused on preservation of, and access to archaeological documents, reports, data sets and images. The most successful digital repositories provide additional value to their users beyond the core mission of preservation. Examples including ArXiv (http://arxiv.org/) or the University of Rochester's digital repository (https://urresearch.rochester.edu) are successful because of additional factors such as reputation or community. For tDAR, the additional value is created through research tools developed on top of the repository. These tools aim to promote new synthetic and comparative research using the data sets stored within the repository.

tDAR's architecture includes three architectural components, a backend preservation repository based on the California Digital Library's Micro-Services model, an interactive web interface, and a research

platform. Metadata is stored within tDAR using a extension of the Library of Congress MODS schema, modified to add archaeologically significant metadata. This includes descriptive metadata about the site, location, culture, materials found, among other attributes. Data sets ingested into tDAR function differently from data sets in a traditional repository. Once a data set has been uploaded, users are guided through the process of documenting their data set within the system. This process is designed to focus on identifying non-machine discernible information such as, whether numeric data represents a measurement or count and translating coded values or lookup tables into human-readable values. Once complete, tDAR's additional features provide unique opportunities to compare, contrast, and analyze data within the system.

A significant challenge for many disciplines is the ability to perform synthetic research. Data from archaeological excavations commonly include a mixture of standardized observational data such as Munsell codes to record sediment color and GPS/ GIS readings are combined with more qualitative assessments about artifact types, or the amount of "burning" on faunal elements. Within the context of a specific site, this is easily reconcilable --as the team develops a common understanding of these terms. However, utilizing these classifications outside of a given site, region, or community of archaeologists, can be challenging. Certain data may lend itself to the application of universal classification schemes --including data that is either more scientific or is derived from a well-documented period. However, more qualitative data may not be as easily mapped to a universal classification model --as definitions of terms will vary between archaeologists or over time. Instead, contributors may provide, or develop a unique classification scheme (ontology) to describe their data. These two approaches represent well-tread road within both research and practice, with distinct benefits to each side. However, to perform any useful comparison, a mapping must be developed.

tDAR does not force users to map data to universal data models or classification schema. Instead, the application has developed a different approach -- maintain the data in its original, and capturing the intent of the archaeologist. The application was been developed with reference ontologies available for certain data elements including faunal species data among others. tDAR enables users to create additional ontologies within the repository, or upload existing one using the OWL format. Once a column of data is associated with an ontology, users are presented with straightforward tools to map the unique data values to terms within the ontology. We believe that this process serves a number of purposes, not only does it maintain and represent the data as it was collected, but it provides opportunities for collaboration and communication within the discipline as archaeologists share data, and discuss intents.

Once data has been mapped, the application guides the user through the data integration process of selecting data sets, identifying columns to compare, fine tuning any mapping issues, and producing the new combined data set. In an analog context, or outside of tDAR, this process can be time-consuming for one data set, and overwhelming for multiple. Within tDAR, what would have been a complex process taking days or weeks when performed manually becomes much more fluid, taking hours. With the technology performing much of the heavy lifting, it leaves the archaeologist to focus on the specific questions and details of their research.

While tDAR is still developing, tDAR's data integration has already enabled Archaeozoologists to ask novel questions about the cultural and ecological circumstances under which species are overhunted or subsistence strategies change. It is our hope that tDAR's core values of access, preservation, and integration will enable us to ask, understand, and evaluate new questions and ideas otherwise impossible within the field of archaeology.

On the Meaning of the Term 'text' in Digital Humanities

Caton, Paul
pncaton@gmail.com
Centre for Computing in the Humanities, King's College London

In digital humanities the word "text" (in both mass and count noun senses)[1] occurs ubiquitously; familiar uses include text encoding, full text search, there are six texts online, we are using the text of the first edition. Typically the word is not defined in the specific context of its use, nor is there an overarching definition or description so widely accepted that it is taken as given at all times. However, our various uses of "text" show we have a priori assumptions about the nature and scope of its reference. But what are those assumptions? are they justified? do they collectively define "(a) text" for us?[2]

In this preliminary investigation I approach from the outside in. I take a number of marginal cases and of each one ask "is this a *text*?" - because any attempt to answer that question must draw out the assumptions that underlie our common usages. I focus on the count noun sense because discrete entities with boundaries ought to be easier to recognize. In our professional lives we talk about "texts" all the time: oughtn't we to know one when we see one?

The marginal cases I discuss are an encrypted message (Figure 1), a sigil[3] created by English occultist Austin Osman Spare (Figure 2), a minimal unit (Figure 3), and a poster with quoted words on it (Figure 4).

WKHTX LFNEU RZQIR AMXPS
VRYHU WKHOD CBGRJ

Figure 1 - encrypted message[4]

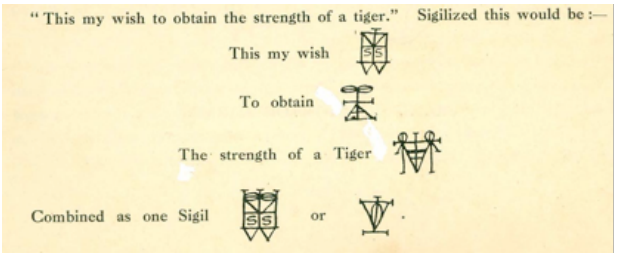


Figure 2 - creation of a sigil from a message string[5]

Figure 3 - minimal unit.

Figure 4 - Second World War poster[6]

2. Summary of Discussion

I suggest the following are core assumptions underlying our collective use of "a text":

- **representation of language**: for any non-metaphorical use we think that language must be involved. Unlike a painting or a piece of music, which seem to affect us unmediated by language, for something to be "a text" it must resolve to language in our heads, even if what we see does not *directly* represent language. Thus we see Milton Glaser's famous logo "I [heart symbol] NY" and in our heads hear the words "I love New York", because the particular symbol-word association is so common that it resolves almost by default - especially given the linguistic context in which the heart symbol occurs. The encrypted message in Figure 1 is all linguistic symbols, but not *directly* interpretable as any language. Resolution to language depends upon knowing how to decipher the symbol sequence - though I suggest that as creatures of language even if we do not know the cipher (and so the sequence remains impenetrable to us) we think it likely that the sequence we see is a reversible transformation of a comprehensible linguistic sequence. We accord it an honorary status as "a text" whose lack of recognition is due to our ignorance, and not to its being something other than "a text". But Figure 2 is a different matter. Unlike the product of the cipher transformation - an incomprehensible string of what are recognisably linguistic symbols - sigilization transforms a comprehensible sequence of linguistic symbols into an almost purely graphic image. I suggest that seeing the final sigil without knowing its origin, we would not even accord it honorary status as "a text", because the deletions, substitutions, and spatial reconfigurations make it almost impossible to resolve back into language - there are so few clues that it started out as language in the first place. On the other hand, the linguistic message has not been replaced by a figurative image, in the way that a photograph of an emaciated child might replace the symbol sequence "children are starving" - indeed mimetic representation of the desire conveyed by the communication would not be to the purpose, "[t]he idea being," writes Spare, "to obtain a simple form which can be easily visualised at will, and has not too much pictorial relation to the desire." (50) For the person who creates the sigil the message is still completely present, implying that for them at least the sigil *is* "a text".
- **communication**: in the normative case for "a text" we assume that a linguistic symbol sequence has been created to communicate, which is the primary function of such sequences. We assume the sequence forms a message (or, in the case of a fragment, would form a message if the entire sequence were present). We have such a propensity to find a message, to make sense out of a sequence, that we will try to establish "a text" even in the least promising cases. Because the glyph shown in Figure 3 represents a character that (in addition to being a letter) is also a lexical item in English it triggers that response, but because the lexical item is supposed to function as a determiner yet here determines nothing it gives us no semantic purchase. Compare this to Figure 3b:

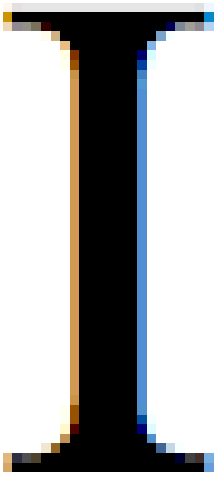


Figure 3b

This is another glyph that represents a character that is both letter *and* lexical item, and here in majuscule form as proper to the lexical item. Because pronouns carry more semantics than indefinite articles it gives us more 'traction'; I suggest that we would rank Figure 3b as *closer* to being "a text" than Figure 3, even if we wouldn't commit to saying that it *is* "a text".

- **completeness**: the completeness of the message embedded in the symbol sequence depends entirely upon the context. What is merely part of "a text" in one context can stand alone in another context, as the excerpt from Churchill's speech does in Figure

4. If we say that the poster in Figure 4 contains "a text", though, does that text contain the words "The Prime Minister" - or are we looking at two texts, one (just the quote) embedded in another (the whole linguistic symbol sequence)? When we hold a paperback book - an edition of *Moby Dick*, for example - how many texts are represented in that physical object?

3. Some Preliminary Conclusions

There is a type of thing called "text" which is a symbol or sequence of symbols that either directly represents language or can be resolved back into language by reversing an earlier, non-arbitrary transformation. In this mass noun sense, text exists, is independent of context, and independent of individual interpretation or experience. However, while text in the mass noun sense must be what makes up text in a count noun sense, there is no such thing as "a text" that is independent of context or of individual experience and interpretation. No linguistic symbol sequence is naturally "organic" or "unitary" (the adjectives used by the TEI Guidelines), though any complex sequence will have structural features that offer themselves as convenient boundaries. Nevertheless these boundaries are always artificial, as much recent work on genetic editions has shown.[7] Being "a text" is a status we give some text in a particular context and at our choosing. In this sense "a text" is, as Renear and Dubin say in a somewhat similar context, "a matter of contingent social/linguistic circumstances" (2007 p.8) and is thus - as they similarly concluded about three of the four FRBR Group 1 entity types - not a type but a role. In other words I suggest that being "a text" is not what Guarino and Welty would term a *rigid property* of any instance of text in its mass noun sense.[8] A good deal of ontological work needs to be done, however, before this can be asserted with confidence.

References

Caton, Paul (1999). 'Using <TEXT> in TEI Markup'. *ALLC/ACH conference.* Virginia, June 1999.

Caton, Paul, and INKE Research Group (2010). 'No representation without taxonomies: Specifying key terms in digital humanities'. *Digital Humanities 2010.* London, July 2010.

DeRose, Steven J., David Durand, Elli Mylonas, and Allen H. Renear (1990). 'What is text, really?'. *Journal of Computing in Higher Education.* 1 (2): 3-26.

Frater U. D. (1991). *Practical Sigil Magic: Creating Personal Symbols for Success. [Trans. Ingrid Fischer.].* St. Paul, MN: Llewellyn Publications.

Guarino, Nicola, and Christopher Welty (2001). 'Supporting ontological analysis of taxonomic relationships'. *Data and Knowledge Engineering.* 39: 51-74.

Rehbein, Malte (2009). 'Reconstructing the textual evolution of a medieval manuscript'. *Literary and Linguistic Computing.* 24 (3): 319-327'.

Renear, Allen, David Durand, and Elli Mylonas (1996). 'Refining our notion of what text really is'. *Research in Humanities Computing [edited by Nancy Ide and Susan Hockey].* Oxford: Oxford University Press.

Renear, Allen H. (corresponding author), and David Dubin (2007). 'Three of the four FRBR Group 1 entity types are roles, not types.'. *Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology (ASIST) [In Grove, Andrew, Ed.].* Milwaukee, WI (US).

Spare, Austin Osman (1913). *The Book of Pleasure (Self-Love).* London: Co-operative Printing Society Limited.

TEI Consortium (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange [Lou Burnard and Syd Bauman, eds.].*

Tonra, Justin (2009). 'Textual studies and the TEI: Encoding Thomas Moore's 'Lalla Rookh''. *Jahrbuch für Computerphilologie.* 10 (2009): 25-36.

Notes

1. By the end of the paper I hope the relation between the mass and the count senses of the noun "text" will be clear. That "text" has a count noun sense is embodied in the Text Encoding Initiative Guidelines (2008 passim).
2. Text encoding models have of course benefited greatly from the "ordered hierarchy of content objects" definition proposed in DeRose et al 1990 and its subsequent refinements such as in Renear et al 1996. Caton 1999 discusses the relation between the TEI element <text> and the concept of "a text". Caton and the INKE Research Group 2010 argues for greater precision in the use of core digital humanities terms such as "text".
3. Sigils in general are magical symbols, and as used by Austin Osman Spare "are developed by fusion and stylization of letters" (Frater, U. D. 1991, 7). The letters come from a sentence that expresses a particular desire of the magical practitioner.
4. Figure 1 is adapted from http://en.wikipedia.org/wiki/Caesar_cipher. Retrieved 28/10/2010.
5. Figure 2 is from Spare 1913, page 50.

6. Figure 4 is from http://en.wikipedia.org/wiki/Never_was_so_much_owed_by_so_many_to_so_few. Retrieved 28/10/2010.
7. Particularly interesting examples are the work of Malte Rehbein on a medieval German town record book (2009), and of Justin Tonra on Thomas Moore's long poem "Lalla Rookh" (2009). In each case the multiplicity of symbol sequences that are candidates for being "a text" is striking.
8. Guarino and Welty define a rigid property as "a property that is essential to all its instances" where by "essential" they follow Lowe in saying that "an essential property of an object … [is one where] the object has that property always and in every possible world" (2001 p.57). An example of a rigid property that they use several times is PERSON: "if x is an instance of PERSON, it must be an instance of PERSON in every possible world" (2001 p.57). They contrast PERSON with STUDENT; STUDENT is a property an entity can have and then not have without the entity changing: the same is not true of PERSON.

Discovering Land Transaction Relations from Land Deeds of Taiwan

Chen, Shih-Pei
gail@turing.csie.ntu.edu.tw
Department of Computer Science, National Taiwan University

Huang, Yu-Ming
ming@turing.csie.ntu.edu.tw
Department of Computer Science, National Taiwan University

Ho, Hou-Ieong
brent@turing.csie.ntu.edu.tw
Department of Computer Science, National Taiwan University

Chen, Ping-Yen
champiyen@turing.csie.ntu.edu.tw
Department of Computer Science, National Taiwan University

Hsiang, Jieh
jhsiang@ntu.edu.tw
Department of Computer Science, National Taiwan University; Research Center for Digital Humanities, National Taiwan University

1. Abstract

Land deeds were the only proof of ownership in pre-1900 Taiwan. They are indispensable for the studies of Taiwan's social, anthropological, and economic evolution. We have built a full-text digital library that contains more than 30,000 land deeds. The deeds in our collection range over 250 years and are collected from over 100 sources. The unprecedented volume and diversity of the sources provide an exciting source of primary documents for historians. But they also pose an interesting challenge: how to tell if two land deeds are related.

In this paper we describe an approach to discover one of the most important relations: successive transactions involving the same property. Our method enabled us to construct over 3,300 such transaction pairs. We also introduce a notion of *land transitivity graph* to capture the transitivity embedded in these transactions. We discovered 2,219 such graphs, the largest of which includes 103 deeds. Some of these

graphs involve land behavior that had never been studied before.

2. Introduction

Until the turn of the 20th century, hand-written land deeds were the only proof of transaction of lands in Taiwan. Such a deed may involve activities such as selling/buying, lending of land to smaller farmers, dividing the land among children or shareholders, and cultivation permits. The deeds usually follow, depending on their types, a typical but not standard format, and are drown up in ad hoc manner. Indeed, even the name of the location may be written in a local convention unfamiliar to the outsiders.

While each land deed may have significance only to its owner, a large collection of them provides a fascinating glimpse into the pre-modern Taiwanese grassroots society. Historians have studied them to investigate the economic activities, community development, and the relationship among the various ethnic groups (Chen, 1997; Ka, 2001; Shih, 2001; Hong, 2005).

In the past few years we have built a full-text digital library of primary historical documents of Taiwan called THDL (*Taiwan History Digital Library*). Among its corpuses is a collection of over 30,000 land deeds, spanning from 1666 to the first decade of the 20th century, and collected from over 100 sources of origin (Hsiang, Chen, Tu, 2009). This collection is unprecedented in terms of volume, time span, geographic distribution, and variety. While THDL presents an exciting source of primary materials for historians, it also poses a challenge: how to find the relationship between two land deeds, or, how to find all the land deeds involving the same piece of land. Although it was customary to hand down earlier deeds to the new owner during the transaction of land, most of these links were broken when the Japanese, during their colonial rule of Taiwan between 1895 and 1945, modernized the land management system (Li, 2004). That is because the officials only recorded the last deed as the proof of ownership but ignored the previous ones. Consequently many of the older deeds were either destroyed or (later) sold as collector's items because they had lost their original value.

In this paper we present a semi-automated method to discover the transaction relations among land deeds. We shall focus on two important relations: *successive transaction pairs* and *allotment agreements*. We further connect the transitive activities on the same piece of land into a concept called *land transitivity graph*, which captures the history of the land over time. The largest such graph that we found has led to a discovery of a new type of land use that had never been observed before.

3. Discovering Land Transaction Relations

We start by describing the two relations among land deeds that our method tries to capture.

Successive transaction pairs: A piece of land could be sold from A to B, then from B to C. In this case there should be two land deeds recording the two transactions. We call them a *successive transaction pair*. Note that the situation could be rather complicated. For instance it could have been B's son who sold it to C. If B divided the land among his descendents, the first selling transaction and the ensuing allotment agreement (see below) also form a successive transaction pair.

Allotment agreements: An allotment agreement is a deed that records how a land is divided among the owner's descendants or among the shareholders. In both cases the usual practice is to first divide the land into several parts, then to have each participant drawing from the lot. Once the decision is agreed upon, an agreement is written, and several copies are made and given to each person involved. In the case of division among shareholders, the allotment agreements should be preceded by a *cultivation permit*, a permission from the government to allow a group of people to cultivate the land. In this case, the cultivation permit and the ensuing allotment agreement also form a successive transaction pair.

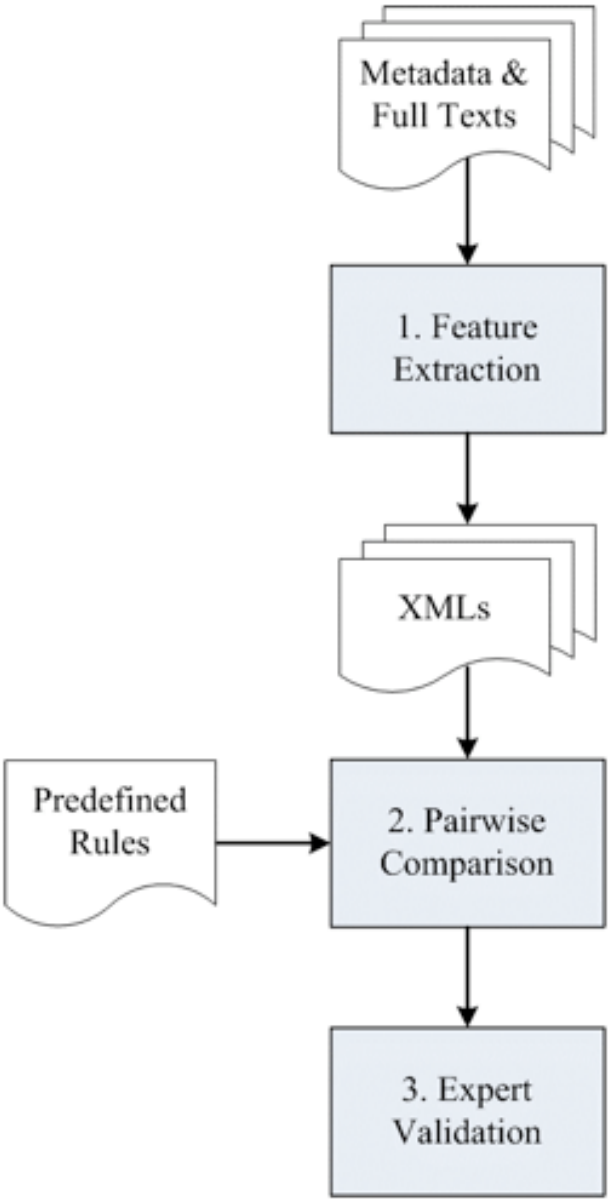


Fig. 1 The process for discovering land transaction relations

To tackle this problem of finding successive transaction pairs, we developed a 3-step semi-automatic process (Fig. 1). We first used text processing technology to extract features of each land deed from its metadata and full text. Such features include the transaction type, the general location of the land and the *four reaches* (boundaries identifying the land via some obscure way such as "bordering Lee's house on the south," "a large camphor tree on the west," etc), the names of the people involved in the transaction and their roles (seller, buyer, scrivener), description of the source of the land (how and when the current owner obtained it), the size, the price, and the amount of taxes paid (Lu, 2008; Huang, 2009). Fig. 2 is an example of a typical land deed. We designed an XML format to hold this information (Fig. 3). Second, we defined rules to identify deeds that may be related. Fig. 4 shows the rules we used for identifying the *successive transaction pairs*. We then wrote a program to compare every pair of land deeds in THDL to see if any pair satisfied the rules. Finally, we give all the pairs produced to human expert to verify.

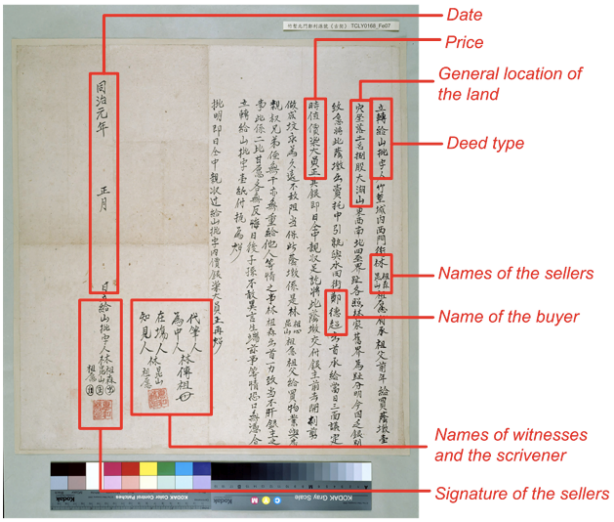
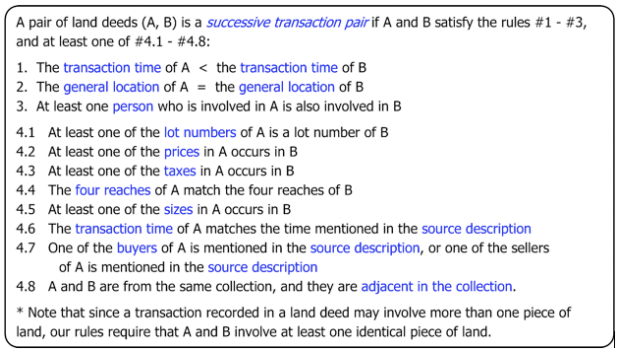


Fig. 2 An example of a typical land deed of Taiwan

```
<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淡沼山林原野ヲ開墾地トシテ整理方認可（臺北廳）</collection>
  <transaction_type>杜賣契</transaction_type>
  <location>一;雙;溪;內;窩;尾;山</location>
  <boundary_E>聖人</boundary_E>
  <boundary_W>嵩脊</boundary_W>
  <boundary_S />
  <boundary_N>余家</boundary_N>
  <seller>何;長;來</seller>
  <buyer>李;宜;周</buyer>
  <time day="18980101" month="189801" year="1898" dynasty="1868" timelevel="year">明治三十一年</time>
  <source_description>水田山園茶樹果子;先祖父置下應得子;過何景奇等山業屋;先問房親人等不欲;受外托中引就與李;買同
    堂議定時值盡;價銀捌百弐拾大員;主前去掌管收租納;終休寸土無留來及;先祖父自置置下應;掛他人財物與及來;主之事此
    乃明實明;二比甘愿並非迫勒;實印契連可單壹紙;茶樹果子竹木屬宇;因乏銀窩用愿賣此</source_description>
  <source_time>1869</source_time>
  <land_size>0.9492</land_size>
  <transaction_price>820</transaction_price>
  <tax>0.483</tax>
  <land_number>四六〇之一</land_number>
</document>
```

Fig. 3 The features of a "selling" type of land deed, stored in XML

We further remark that a criterion that allows certain degree of fuzziness was used when performing matching. This is because the names used in different deeds may sometimes be slightly different even if they are the same place or person (Huang, 2009).


A pair of land deeds (A, B) is a *successive transaction pair* if A and B satisfy the rules #1 - #3, and at least one of #4.1 - #4.8:

1. The transaction time of A < the transaction time of B
2. The general location of A = the general location of B
3. At least one person who is involved in A is also involved in B

4.1 At least one of the lot numbers of A is a lot number of B
4.2 At least one of the prices in A occurs in B
4.3 At least one of the taxes in A occurs in B
4.4 The four reaches of A match the four reaches of B
4.5 At least one of the sizes in A occurs in B
4.6 The transaction time of A matches the time mentioned in the source description
4.7 One of the buyers of A is mentioned in the source description, or one of the sellers of A is mentioned in the source description
4.8 A and B are from the same collection, and they are adjacent in the collection.

* Note that since a transaction recorded in a land deed may involve more than one piece of land, our rules require that A and B involve at least one identical piece of land.

Fig. 4 The rules for identifying successive transaction pairs

The precision rate of the algorithm for successive transaction pairs is 63.9% and that for allotment agreements is 94.4%. We have found 2,409 successive transaction pairs and 878 sets of allotment agreements among the 30,820 land deeds in THDL (Table 1).

Relationship	Relations discovered	Cross sources	From same source but files are not adjacent
Successive transaction pairs (pair)	2,409	119	738
Allotment agreements (set)	878	56	208

Table 1 The result of reconstructing land transaction relations among the land deeds in THDL

Among the former, 358 are cross-generation (A sold to B, and B's descendent sold to C). Some of the pairs/ sets are from different sources (the "cross sources" column in Table 1), and are quite impossible to find manually. Some others are from the same source but are not adjacent to each other in their original order. These are also difficult to identify by hand.

4. Land Transitivity Graphs

When further examining the transaction pairs, an interesting transitive phenomenon emerged. There may be a deed of A selling a piece of land to B, and some years later B divided the land among his sons, then one of them, C, rented it to D to farm. Such transitive activities on the same piece of land could last for decades. By connecting all these transactions into a graph, it may capture the evolution of a property over time.

This is exactly what we did. We call these graphs land transitivity graphs. Using the relations we discovered early, we came up with 2,219 such graphs. The result is listed in Table 2.

Deeds Involved	Number of Graphs
103	1
65	1
36	2
23	1
15-19	7
10-14	27
6-9	112
5	105
4	218
3	412
2	1,333
Total	***2,219***

Table 2 The land transitivity graphs constructed among the land deeds in THDL

Fig. 5, the third largest graph, contains 36 deeds, dating from 1850 to 1910. The head of the family, Liao Jiafu (廖佳福), was among the shareholders who received a cultivation permit from the Qing government, and obtained this piece of land through allotment in 1850 (the first deed).

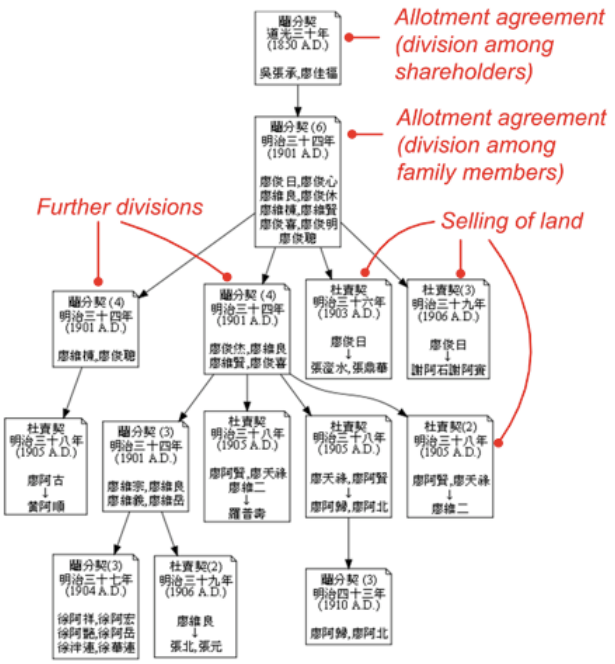


Fig. 5 The 3rd large graph, containing 36 deeds

Liao farmed the land for 50 years and divided it among his descendents in 1901 (the second deed). The rest of the deeds described the various activities such as further divisions or selling in the next 10 years. By 1906, only 2 of the 8 divided pieces of land remained in the Liao family.

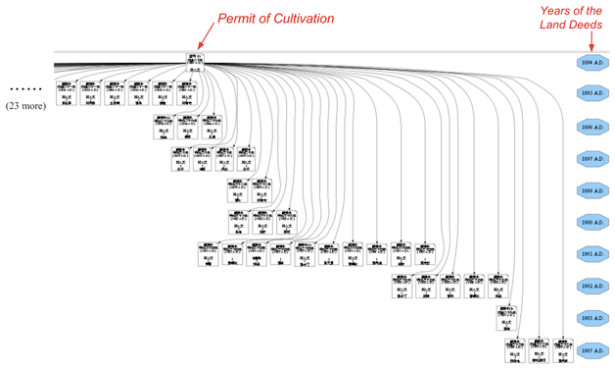


Fig. 6 The largest graph, containing 103 deeds

Fig. 6 is the largest land transitivity graph with 103 deeds. Tu, a historian, studied this graph and discovered that the deeds involved demonstrated a unique case of land use that had never been studied before (Tu, 2010). It is unlikely for human to notice this possibility without the computer-generated transitivity graph.

To help historians take advantage of these graphs, we developed an integrated environment to analyze the information embedded in each graph (Fig. 7). In addition to the graph itself and its zoomable navigation facility, we also added tag cloud, chronological distribution, and a location map.

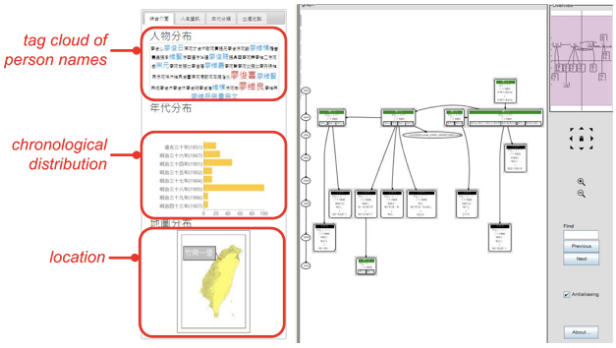


Fig. 7 The integration environment for land transitivity graphs

5. Concluding Remarks

Land deed research has been an important topic among historians of pre-1900 Taiwan. In this paper, we presented a method to discover the transaction relations among the 30,820 land deeds in THDL, the largest existing full-text database of land deeds. Our method discovered 2,049 successive transaction pairs and 878 sets of allotment agreements. They, in turn, are transformed into 2,219 land transitivity graphs, each of which describes the transaction evolution of a piece of land. One such graph has already led to the discovery of a unique pattern of land development that had not been studied before (Tu, 2010). We feel that our work demonstrates how IT tools can be used to help historians conduct research that could not be done otherwise.

References

Chen, C. K. (1997). *Taiwan's aboriginal proprietary rights in the Ch'ing period: Bureaucracy, Han tenants and the transformation of property rights of the Anli Tribe, 1700-1895.* Taipei: Academia Sinica.

Hong, L.W. (2005). *A study of aboriginal contractual behavior and the relationship between aborigines and Han immigrants in west-central Taiwan.* : Taichung County Cultural Center V. 1. .

Hsiang, J., Chen, S. P., Tu, H. C. (2009). 'On building a full-text digital library of land deeds of Taiwan'. *Digital Humanities 2009 Conference.* Maryland, June 22-25, 2009, pp. 85-90.

Huang, Y. M. (2009). *On reconstructing relationships among Taiwanese land deeds. Master thesis.* Taipei, Taiwan: National Taiwan University.

Ka, C. M. (2001). *The aborigine landlord: Ethnic politics and aborigine land rights in Qing Taiwan.* Taipei: Academia Sinica.

Li, W. L. (2004). 'Land deeds and land administrative documents—Interpreting the Archives of the Japanese Taiwan Governor-Generals'. *Taiwanese History Research.* 11 (2): 221-240.

Lu, C. C. (2008). *Automated Classification of Taiwanese Land Deeds. Master thesis.* Taipei, Taiwan: Taiwan University.

Shih, T. F. (2001). *Local society in Qing Taiwan.* Hsinchu: Cultural Affairs Bureau.

Tu, F. E. (2010). 'Environmental Change, Land development and Dispute over Property Rights in southern Taiwan (1890-1920)'. *The Sixth Conference of Taiwan Colonial Government Archives.* Taiwan Historica.

Names in Novels: an Experiment in Computational Stylistics

van Dalen-Oskam, Karina
karina.van.dalen@huygensinstituut.knaw.nl
Huygens Institute for the History of the Netherlands - KNAW, The Hague

The growing amount of digital texts and tools is slowly but definitely changing the way literary scholars design their research. This paper will describe the effects for one research topic: ‘literary onomastics’, the study of the usage and functions of names in literary texts. Research in literary onomastics usually is qualitative in nature and focuses on 'significant' names in literary texts. No quantitative or comparative studies have been published yet. Several researchers, however, have pointed out that names can only be called significant if they are studied in the context of all the names - the so-called 'onymic landscape' - in a text, oeuvre, genre etc. (e.g. Sobanski 1998). This question is comparative by nature and implies the wish for a more quantitative, computer-assisted approach.

As to name usage, first we have to find out what is normal. We want to know how many name forms usually occur in a novel, and how many of these are personal names, place names, and other names. As to name functions, a useful disctinction is between plot internal and plot external names. Plot internal names refer to characters, places or other entitities which only ´exist´ in the fiction of the story. Plot external names refer to persons, places or other entities which are known to exist or to have existed in the real world. Most place names are plot external, referring to real countries, cities, streets, etc. and thus have a reality enhancing function. In fantasy novels, however, place names are usually fictional and thus plot internal, enhancing the unreality, the fantasy of the story. Plot external personal names seem to function as characterizations of the fictional characters, describing e.g. their political or cultural preferences.

It is expected that different authors, genres, time periods or even languages apply these different types and functions in different ways, showing different trends which we want to discover in what we like to call comparative literary onomastics (Van Dalen-Oskam 2005, 2006). The ultimate aim of the research is to develop a method to compare the name usage and functions on as large a scale as possible, explicitly also across languages. So what is needed to do that?

The amount of names in a text can be expressed in the percentage of the total amount of tokens in the text. For that, we need digital texts of fair to good quality. To tag all the names, we need named entity recognition and classification (NERC) tools. Different forms of the same name (e.g. *Patrick* and *Patrick's*) need to be grouped by a ***lemma*** (PATRICK in this case). Different name forms for the same person or place need to get the same ***identification*** (e.g. the name *Alfred* and the name *Issendorf* both belong to the character identified as *ALFRED ISSENDORF*). To find out whether we can compare the resulting percentages across languages, we focused on a corpus of modern Dutch and English novels and their translations into the other language. We found we have to include two other levels of aggregation: ***mentions*** and ***name tokens***. Mentions are occurrences of a name irrespective of the number of tokens used. So several name tokens can be used in one mention. This distinction is necessary because different languages have different tokenization rules. The Dutch personal name *Gerrit-Jan*, e.g., with a hyphen and therefore counted as one token, is translated in English as *Gerrit Jan*, resulting in two tokens. Our corpus consists of 22 Dutch and 22 English novels, added with the translation into the other language of ten in each group.

Comparative research can only be done when many scholars collaborate. We will have to make sure that all those scholars encode their texts in the same way, considering the same tokens as names. This may sound easy, but it is not. Even name theorists have different definitions of what a name is (Van Langendonck 2007). Guidelines had to be set. We decided to limit the tagging to the 'prototypical' names, so those names that are considered names by readers in general. Something is a name if it refers to a unique person, place, or object. So we excluded currencies, days of the week, months, etc. For cases leading to discussion we defined additional rules.

We found that not many modern novels are digitally available yet. Furthermore, NERC tools proved to be not good enough yet and are using partly different categories than our research needs (Sekine and Ranchod 2009). This meant much more manual work than we expected. The tagging and the statistics are now done by means of several perl scripts written by André van Dalen and an ingenious Excelsheet with macros designed by David Hoover. These tools can be seen as a prototype for the webservices we need for this type of comparative stylistic research.

In Figure 1, the percentage of name tokens in a set of novels is shown. The selection of novels in this graph consist of sixteen novels of which fifteen are Dutch and one is (American-)English (*Of the Farm*), added with the English translation of one of the Dutch novels (*The Twin*, translation of *Boven is het stil*).

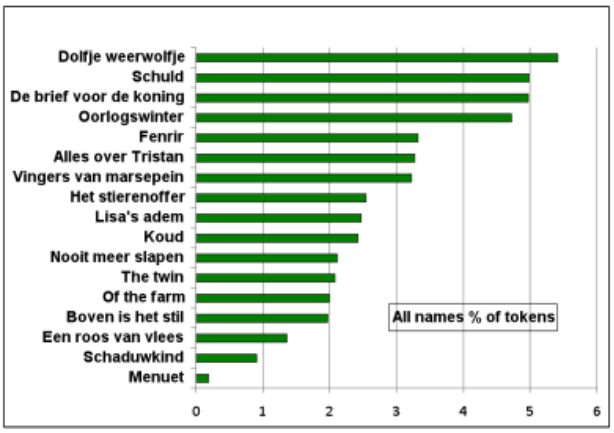


Fig. 1

The graph gives as a first impression that we can expect around 2-3 percent of the tokens in a novel to be names. For this small selection, it is noteworthy that the four novels with the highest percentages (around 5 percent) are all books for children or young adults. Furthermore, the English novels do not show up in extreme positions compared to the other ones.

Figure 2 gives more insight in the role of plot internal versus plot external names in the same set of novels. The novels are presented in the same order as in Fig. 1. We can see that the four books for children/young adults are still exeptional in the percentage of plot internal names. We also find that three of them have a rather small amount of plot external name tokens.

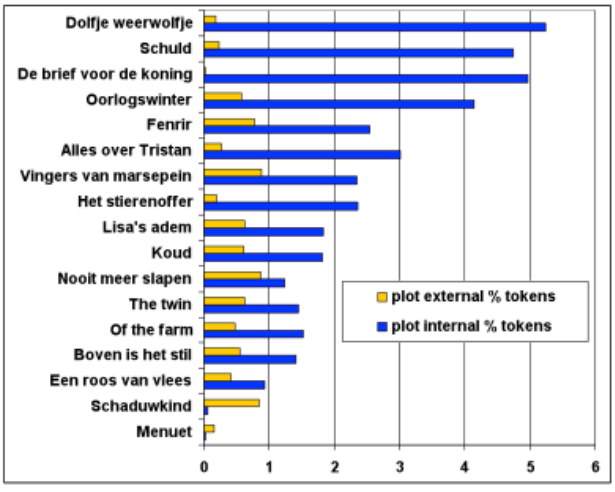


Fig. 2

The paper will present the case of names in Gerbrand Bakker's novel *Boven is het stil / The Twin*. Dutch readers have the impression that the novel does not

contain many names, while readers of the English translation have the opposite impression and point out that especially geographical names abound. The comparative computational approach shows that both reader groups are correct *and* wrong at the same time. The novel has around 2% of name tokens, which is at the lower end of what seems to be normal. But in the amount of *different* names (lemmas), geographical names (place names) do have a special role when we look at the ratio between personal names and place names (Figure 3). The trend here is that a novel contains more different personal names than place names. In only two cases (*Boven is het stil* and its translation *The Twin*, and *Fenrir*) a novel has more different place names. This suggests that place names have an extra stylistic function in these two novels. It will be shown in which ways place names function in the Bakker novel and how this analysis enriches the understanding of the novel *and* of the stylistic possibilities of place names for an author.

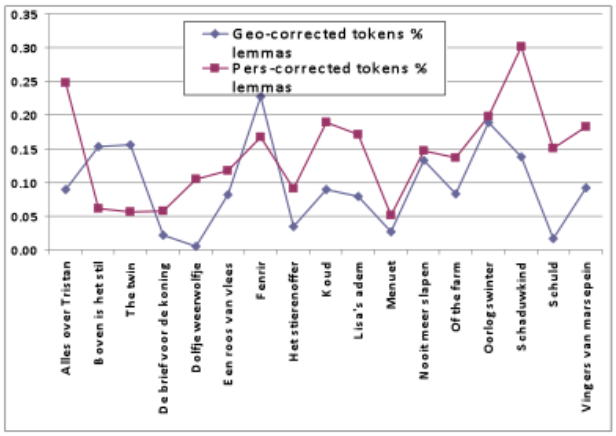


Fig. 3

We could only show a small part of all interesting observations to be made about the usage of names in a corpus of modern Dutch and English novels. But these first results make us anxious for more, in the expectation that this approach may lead to an acceptable method for a.o. across language comparison of stylistic elements.

We conclude that the preliminary results are sufficiently interesting to go into the stylistic analysis of name usage and functions in novels more deeply. Names also seem promising stylistic elements for a comparison across languages. The currently available tools which could be expected to be helpful for this type of research, proved to be insufficient. We therefore plan to develop a set of interrelated webservices which will assist the scholar in the recognition, categorization, further tagging, and statistical analysis of names in novels.

References

Sekine, S. and Ranchod, E. (2009). *Named Entities. Recognition, classification and use.* Amsterdam/ Philadelphia: John Benjamins.

Sobanski, I. (1998). 'The Onymic Landscape of G.K. Chesterton's Detective Stories'. *Proceedings of the XIXth International Congress of Onomastic Sciences.* Aberdeen, 1996, pp. 373-378.

Van Dalen-Oskam, K. (2005). 'Vergleichende literarische Onomastik'. *Namenforschung morgen: Ideen, Perspektiven, Visionen.* Brendler, A. und S. Brendler (ed.). Hamburg: Baar, pp. 183-191. `http://www.huygensinstituut.knaw.nl/wp-content/bestanden/pdf_vandalenoskam_2005_Comparative_Literary_Onomastics.pdf`.

Van Dalen-Oskam, K. (2006). 'Mapping the Onymic Landscape'. *Il nome nel testo. Rivista internazionale di onomastica letteraria VIII; Atti del XXII Congresso Internatzionale di Scienze Onomastiche.* Pisa, 28 agosto – 4 settembre 2005, pp. 93-103.

Van Langendonck, W. (2007). *Theory and typology of proper names.* Berlin / New York : Mouton de Gruyter.

Victorian Women Writers Project Revived: A Case Study in Sustainability

Dalmau, Michelle
mdalmau@indiana.edu
Indiana University Digital Library Program

Courtney, Angela
ancourtn@indiana.edu
Indiana University Libraries

The *Victorian Women Writers Project* (*VWWP*, `http://www.dlib.indiana.edu/collections/vwwp/`) began in 1995 at Indiana University under the editorial leadership of Perry Willett and was celebrated early on for exposing lesser-known British women writers of the 19th century. The *VWWP*'s original focus on poetry was meant to complement *The English Poetry Full-Text Database*, but soon Willett acknowledged the variety of genres in which women of that period were writing – novels, children's books, political pamphlets, religious tracts. The collection expanded to include genres beyond poetry, and continued active development from 1995 until roughly 2000 at which point the corpus reached approximately two hundred texts.

These nearly two hundred texts comprise only a small fraction of Victorian women's writing. Encouraged by renewed interest among Indiana University's English faculty and graduate students, the Indiana University Libraries and the English Department are exploring ways to reinvigorate the project. The real challenge lies in the project's past and present susceptibility to graceful degradation, which can be defined as stagnation or "deterioration of a system in such a manner that it continues to operate, but provides a reduced level of service rather than failing completely" ("Graceful degradation"). This has been a recurring topic in the digital humanities community as evidenced by a recently published article cluster, "Done," in the *Digital Humanities Quarterly* (Spring 2009, v 3 n 2) and research by Bethany Nowviskie and Dot Poter on the very topic:

"Decline is a pressing issue for digital scholarship because of the tendency of our projects to be open ended. One could argue that digital projects are, by nature, in a continual state of transition or decline. What happens when the funding runs out, or the original project staff move on or are replaced? What happens when intellectual property rests with a collaborator or an institution that does not wish to continue the work? How, individually and as a community, do we weather changes in technology, the patterns of academic research, the vagaries of our sponsoring institutions?" ("Graceful Degradation: Managing Digital Projects in Times of Transition and Decline").

The *Orlando Project* is a text-based resource containing primary texts, archival documents, biographies, chronologies and bibliographies. Despite the collection's extensive size, Brown et al. reveal their sustainability challenges in an article published as part of the "Done" cluster:

Lack of people, time, or funding has consigned more than one project involuntarily to becoming a static tribute to its former activity. The reasons for this include people moving on, intellectually or institutionally, without taking their projects along with them, or people using electronic media to disseminate without particularly desiring to exploit their potential for continual updating, but even where the will to continue persists, inadequate funding mechanisms for sustainability may make it impossible. This is a shame, since, as we have argued here in the case of *Orlando* and many other digital publications not only does there remain the potential to enrich the contents, but the first iteration often merely begins to tap the potential of the project's data architecture and potential for interface development ("Published Yet Never Done").

Unfortunately and to no surprise, the *VWWP*, quite modest compared to its *Orlando Project* counterpart, has also suffered from nearly all the challenges highlighted by Nowviskie, Porter, and Brown.

In an effort to combat this "darker side of project management," a framework for continual project support is being explored that reaches beyond any one individual or department (Nowviskie and Porter). At the crux of this framework is digital humanities-focused curriculum-building for the English department in partnership with the library, with a concentration on scholarly encoding and textual editing, working with English faculty, librarians, and technologists. The goals are to leverage domain expertise in the English Department; integrate the *VWWP* as a core research and teaching tool in the English curriculum; develop TEI and text encoding expertise in faculty and students; and through coursework, internships, and other opportunities, encourage English literature students — graduates and undergraduates—to continually

contribute new content to the *VWWP*. Tools and workflows, such as robust encoding guidelines, quality control assessment, etc. will be provided to ensure proper markup, and the *VWWP* editorial board will additionally vet course output before submission to the project.

Through our newly offered graduate English course (L501, Digital Humanities Practicum), an eager and curious group of students learned not only encoding skills but also began to develop the collaborative practices pervasive in the digital humanities. As part of our talk, we plan to explore whether cultivating "markup skills" are sufficient enough in establishing a digital humanities curriculum (Rockwell) and whether "majoring in English" today means the curriculum should include awareness of the possibilities that arise for new scholarship when technology is applied to literary studies (Lanham). Certainly Indiana University is not breaking new ground or alone in this endeavor, but the literature is scarce is terms of understanding successes of graduate level digital humanities curricula situated in an English or any other humanities department. As Diane Zorich reports in her recent review of digital humanities centers, "A Survey of Digital Humanities Centers in the United States," archives such as the Willa Cather and Walt Whitman Archives are precisely leveraged for teaching and learning, and this reporting is promising for the *Victorian Women Writers Project* as a project reconceived to meet both teaching and research needs in a classroom setting (19).

Currently, the *VWWP* is a standard e-text project, although current plans call for phased, modular development that will eventually include now commonplace "Web 2.0/3.0" functionalities. By garnering institutional commitment (at the risk of wavering priorities) across multiple departments (thereby minimizing risk), we hope to achieve the following:

- Encourage English department buy-in and continual collaboration by updating the current state of the *VWWP*'s functionality and modernizing the look-and-feel (eliminate "first impressions" syndrome discussed by Brown et al.)
- Establish a sustainable scholarly encoding infrastructure based in the English department curriculum
- Provide a consistent mechanism (e.g., coursework output) for critical content to accompany the encoded texts
- Facilitate connections between other DLP-supported Victorian projects like the *Swinburne Project* and the *Victorian Studies Bibliography*
- Evolve the project's encoding guidelines, inclusion of critical contextual materials, and advanced functionality (e.g., visualizations, textual analysis tools, blog integration, etc.) so that the *VWWP* becomes a dependable, growing and relevant online resource that can be adopted as a pedagogic and research tool for Victorian scholars

Our talk will introduce the *Victorian Women Writers Project*, explore curriculum-building strategies; and propose ways in which faculty and students can reliably and perpetually contribute to the *VWWP*.

References

Brown, Susan (2009). 'Published Yet Never Done: The Tension Between Projection and Completion in Digital Humanities Research.'. *Digital Humanities Quarterly.* Web 3.2. `http://digitalhumanities.org/dhq/vol/3/2/000040/000040.html` (accessed 12 April 2010).

(23 Aug. 1996). 'Graceful Degradation'. *Federal Standard 103C: Glossary of Telecommunications Terms.* Web: Institute for Telecommunication Sciences. `http://www.its.bldrdoc.gov/fs-1037/dir-017/_2479.htm` (accessed 12 April 2010).

Lanham, Richard (1989). 'The Electronic Word: Literary Study and the Digital Revolution'. *New Literary History.* Web, pp. 265-290. `http://www.jstor.org/stable/469101` (accessed 30 October 2010).

Nowviskie, Bethany (10 July 2009). 'Graceful Degradation'. Web. `http://nowviskie.org/2009/graceful-degradation/` (accessed 10 April 2010).

Porter, Dot, Bethany Nowviskie (7-9 September 2009). 'Graceful Degradation: Managing Digital Projects in Times of Transition and Decline'. *Digital Resources for the Humanities and Arts.* Belfast, Ireland.

Rockwell, Geoffrey (2003). 'A Graduate Education in Humanities Computing'. *Computers and the Humanities.* Web 37.3: 243-244. `http://www.jstor.org/stable/30204899` (accessed 30 October 2010).

Willett, Perry (1996). 'The Victorian Women Writers Project: The Library as a Creator and Publisher of Electronic Texts'. *The Public-Access Computer Systems Review.* Web 7.6. `http://epress.lib.uh.edu/pr/v7/n6/will7n6.html` (accessed 10 April 2010).

Zorich, Diane (2008). *“A Survey of Digital Humanities Centers in the United States.”*. Web: Council on Library and Information Services (): . http://www.clir.org/pubs/reports/pub143/pub143.pdf (accessed 30 October 2010).

Reusability of Literary Corpora: the "Montaigne at work" Project

Marie-Luce Demonet
marie-luce.demonet@univ-tours.fr
Centre d'Etudes Supérieures de la Renaissance, UMR-CNRS , Université Francois-Rabelais, Tours, France

1. Introduction

I shall examine to what extent the BVH (Virtual Humanistic Libraries) project on Montaigne could be considered not simply another electronic edition, but also a component of a digital humanities infrastructure, observing the keywords of an integrated search: reliability, sustainability, dissemination, and above all, reusability. Is a project about Montaigne's work compatible with the "genericity" required for an undertaking that concerns a wide community?

2. "Montaigne at Work" in the BVH Website

The Bibliothèques Virtuelles Humanistes (http://www.bvh.univ-tours.fr) offers facsimiles (jpeg and light pdf) of books or manuscripts, extracted graphics with their indexing systems, and a textual database called Epistemon. It offers two types of digital surrogates of the book: the facsimile and, for some documents written in French, the corresponding transcription without modifications.

In itself, the idea of digitizing Montaigne's complete works is not original. A "Corpus Montaigne" already exists on CD-Rom but no online version is available at present, and access is limited to the few libraries that could afford it.[1]The "Montaigne project" in Chicago (P. Desan) offers many documents related to Montaigne: it displays the "Villey" edition with every page of the "Bordeaux copy" (Exemplaire de Bordeaux, the so-called "EB"), but the distinction between the three main layers (1580-1582, 1588, 1588-EB) does not comply to the requirements of modern philology and scholarship: cancellations are not visible, and the editor modified the punctuation as well as the spelling.[2]The 1595 (posthumous) edition is already available in HTML format at the mysterious Trismegiste website, but there is no XML encoding, and the spelling is regularized. [3]In our project, all the editions will offer the double display of original/regularized spelling; indexes of names, places, errata, and a basic encoding appropriate for

early printed books and manuscripts. Easy retrieval of both versions, in the format preferred (XML/TEI, HTML, PDF) will be the user's choice, according to the principle of reusability.

As we share our expertise with cultural institutions, we borrow our techniques and methods of digitization of cultural heritage objects, such as rare books collections, from libraries and archive repositories: digitization, metadata organization and catalogs, and database management. Our membership in the European Library (Europeana) helps to understand the difference between a cultural heritage attitude and a research project.[4]The parallel display of the facsimiles and their transcriptions, TEI encoding, tools for scholarly annotations and an accurate query system are not simple challenges to take on: the uniqueness of every work of art, the complexity of the process of writing seems incompatible with the unified view of textual databases usually found in library websites (e.g. Shakespeare at the British Library) or linguistic corpora (Frantext database or ARTFL in Chicago). Scholarly annotation will be minimal, and limited to the accuracy of the transcription, in order to provide a basis for further commentary, encyclopedic information, and glossaries. The very process of building the corpus for online publication is a field of new research in this case, for it combines ergonomic full display and retrieval, complex and relevant extraction procedures, treatment of texts and graphics.

The "Montaigne at work" project aims to support both the reading and the mining of the text, and to render the chronology. Our new data, expertise, and tools will try to fulfill the main goal we have always had of understanding Montaigne's *Essays better*: 1) to offer a *genetic edition*of the "Bordeaux copy," containing several layers of handwritten additions that reveal up to seven moments of writing or re-writing; 2) to give access to what is left of the famous "Librairie de Montaigne". The main corpus (all the editions from 1580 to 1595, with their transcriptions) will be enshrined inside a wider set of later editions of the *Essais* (Marie de Gournay's copy, Rousseau's copy), of several other works of Montaigne (the translation of the *Theologie naturelle*by Raymond Sebond, the *Journal de voyage*), of all his surviving manuscripts (marginalia, letters and Parliament archives), and of facsimiles of about a hundred identified sources (mainly classics, but also books by his contemporaries).

3. Genetic Encoding

The genetic edition of the Bordeaux copy, compatible with the TEI schemas for manuscripts and prints, and the "TEI Renaissance encoding" protocol developed in Tours,[5]raises the question of the relevance of such an undertaking. It must from with a benchmarking of other websites offering open access to digitized works of late Medieval and Early Modern period (Chaucer, Dante, Shakespeare, Cervantes, Descartes, Molière,...). What kind of textual properties do these sites represent? Do they use several models? Exclusive tools? Many literary projects, particularly in France, do not use TEI encoding (Flaubert in Rouen, Montesquieu in Lyons, Stendhal in Grenoble), and scholarly corpora seem to be specific to each author.

Classicists and Medievalists have opened many doors, and they know quite well how to refine an ultra-diplomatic encoding and display. Rendering the writing process requires the adequate edition to feed every hypothesis about the moments of the gesture itself, the "traits de plume" (pen strokes), and the modifications the printing press of the time forced upon the original. Special software designed by our computer science partners (in Tours, Paris, Rouen, La Rochelle) is currently being developed to detect image similarity. Thus, Montaigne's different "hands" could be classified according to time and language, with the expert help of Alain Legros (researcher in Tours, and an expert in Montaigne's handwriting).[6]We need also the clearest visualization of the readable parts, the possibility of displaying either a smoothed text or a page, which represents all the complex arrangements of the words in a spatio-temporal order. No models seem to be directly reusable: ours would take place between the very precise reconstitution of all the spellings of Medieval texts (e.g. the *Actes des Apôtres*project) and the *Madame Bovary*digital edition of manuscripts at the University of Rouen, but with a display system that would look like the Deutsche Text Archive (DTA): the facsimile of the page linked to the HTML text, and to the XML/TEI source, searchable with PhiloLogic (Mark Olsen, University of Chicago) and other NLP tools, with the XTF search engine.[7]All the quotations of the Bordeaux copy will be fully referenced and translated in French.

In Tours, we have already begun the keyboarding and encoding of the main editions (1580-82, 1588, 1595). The genetic edition of the Bordeaux copy will be based on the principles of the ITEM laboratory (École Normale Supérieure, Paris), a leader in genetic analysis, which are compatible with the TEI tagging of the main operations (addition, deletion, inversion,

etc.), according to the latest documentation of the TEI consortium. The COST "Interedition" project (funded by Europe) offers several tools to test (e.g. Collatex for the main editions), and discusses some issues close to our preoccupations, such as the limits of crowdsourcing: we plan to use collaborative annotation by scholars for corrections of errors. [8]

Every layer of text must be retrievable, to avoid incompatibility between the genetic and the generic, and to guarantee reusability to anyone who wishes to process the text (with permission) for other purposes. Ideally, a collaborative edition of Montaigne's *Essais*would blossom out of the accurate transcription of the Bordeaux copy, and/or of the posthumous 1595 edition: the debate is still pending among specialists.

4. Automatic Regularization

We will generate three levels of transcripts:

1. the "quasi-diplomatic" transcript, crucial for the comparison between the typeset and the handwritten passages (the spelling of which has never been thoroughly studied)
2. the "cultural heritage" transcription that regularizes the distinction of I/J and U/V, expands the brevigraphs and normalizes the ends-of-lines, so that the corpus can be processed by the NLP tools and parallel corpora analysis
3. the modernized version, so that powerful search engines can offer accurate results to anybody.

A prototype of I/J U/V normalization tool is already prototyped in Tours and Poitiers, with a set of rules and specific dictionaries; the modernization tool is in progress, and requires another set of rules, and other dictionaries.

The development of these tools benefits one of the two Google awards that the University of Tours obtained in December 2010 for "Full-text retrieval and indexation for Early Modern French documents". New software will process a sentence such as:

Ie veus qu'ō m'y voie en ma façō simple, naturelle & ordinaire, sans estu de & artifice: car c'est moy que ie peins(Montaigne, *Essais*, 1580)

With these spellings, the user who is not a specialist will find only few results in his word or string query because of typographic abbreviations (*façō*for *façon*), obsolete morphology (*veus*for *veux*), and the frequent lack of hyphenation. In modern editions, one will find easily « estude » in the editions following former spellings; but if one looks for « étude » (in modern French), the old spelling will not be offered, and one will miss the variant « estude » in the corpus, where moreover the word is typed without hyphenation.[9]

5. Montaigne's Library

Thus, Montaigne's library itself can be rebuilt through the comprehensive digitization of what remains of the hundred known copies with his signatures and annotations: 33 are preserved at the French National Library, 30 at the Bordeaux Public Library, others in at least 15 other libraries and private collections. Samples of his handwriting will be analyzed and compared to non-attributed manuscripts, in order to confirm or exclude dubious documents.

Such a project will enlarge the knowledge we already have of Montaigne's method of writing, within the context of his favorite readings. If other projects provide data, this one offers also reusable sets of transcriptions, facsimiles and new tools for further analysis.

Notes

1. *Corpus Montaigne*, Paris: Champion-Garnier électronique, 1999.
2. Montaigne project, http://www.lib.uchicago.edu/efts/ARTFL/projects/montaigne/
3. http://www.bribes.org/trismegiste/es1ch03.htm.
4. http://www.europeana.eu/portal/.
5. *Manuel d'encodage TEI-Renaissance*, 2009, http://www.bvh.univ-tours.fr/XML-TEI/index.asp.
6. LEGROS, Alain, *Montaigne manuscrit*, Paris : Garnier, 2010.
7. *Actes des Apôtres*, http://eserve.org.uk/anr/; DTA, http://www.deutschestextarchiv.de/; *Madame Bovary*, http://bovary.univ-rouen.fr/; XTF, http://xtf.cdlib.org/documentation/programming-guide/.
8. http://www.interedition.eu/.
9. Cf. the Old English variation analysis in the York-Helsinki corpora (http://www.helsinki.fi/varieng/CoRD/corpora/YCOE/index.html).

Joanna Baillie's ***Witchcraft***: from Hypermedia Edition to Resonant Responses

Eberle-Sinatra, Michael
michael.eberle.sinatra@umontreal.ca
Université de Montréal, Canada;

Crochunis, Tom C.
TCCroc@ship.edu
Shippenberg University

Sachs, Jon
jsachs@alcor.concordia.ca
Concordia University

This paper will report on the first eighteen months of a 3-year grant funded by the *Fonds québecois de la recherche sur la société et la culture* led by playwright Patrick Leroux (overseeing the creative component) and Michael Eberle-Sinatra (overseeing the academic component). The specific nature of this group project is nestled in the promising dialogue to be established between Romantic literature scholars, a theatre practitioner, and a scholar preoccupied with the pedagogy of Romantic drama using hypermedia as a template and an engaging interface.

When artists teach, they never quite relinquish their initial creative impulse. Historical works, while being taught for their intrinsic value and larger pertinence within a literary context, nevertheless solicit a resonant response. Classroom exercises in both academic and creative courses suggest that many students engage in a similar empathic manner when allowed to prod, question, and interact actively with a studied text.

The "creative" component of this research-creation project with strong pedagogical intent is precisely linked to an artistic response to the source text, Joanna Baillie's *Witchcraft*. In addition to the edited text, its scholarly annotations and commentary, and the filmed Finborough production of the play, we will create workshop situations with actors and students in which the play will be explored in rehearsal prompting us to investigate other manners of staging the work and illustrating, through filmed documentation, the *process of reading a text for performance*. Short video presentations of key creative and interpretive issues will be edited for inclusion in the hypermedia presentation. The actual process, whether filmed or not, will allow the actors and creative faculty to fully immerse themselves in Baillie's world and work in order to fuel their resonant responses to them.

This second creative component, the *resonant response*, will take the form of short theatrical pieces conceived for film. The nuance is essential as the pieces will not be short cinematic films but rather short-filmed theatrical pieces. The emphasis on speech, dramatic action, and relationship to a defined theatrical space will differentiate these pieces from more intimate, image-based cinematic pieces. The resonant responses could be as short as two minutes or as long as ten minutes, in order to fully explore very precise formal issues (a character's speech, the subtext in a given dialogue, what we couldn't stage during the 19th Century but feel we could now). These creative pieces will be developed with Theatre, Creative Writing, and English literature students and faculty.

Existing TEI guidelines for scholarly encoding do not account for the unique relationship between a play script and performance practice and history. Scholarly encoding typically views the structures of texts in relation to the protocols that guide how readers interpret documents. But dramatic scripts require different kinds of reading and, thus, different kinds of encoding. Performance-informed inquiry into play texts depends on a reader's ability to think about the range of possibilities—both historically distant and contemporary—for theatricalization of a line of dialogue, a bit of physical action, or a visual space.

Additional historical materials on the theatre and culture of Baillie's era will be provided by team members. For our hypermedia resource to organize multi-media materials in ways that will help students in literature classes to use the hypermedia edition of the play, we will need to develop innovative customizations of TEI encoding guidelines. Discovering how best to support a student reader's work with a historically unfamiliar dramatic work provides an important test case for existing guidelines for XML encoding of drama.

This project will take an innovative approach in several senses. It will use hypermedia to try to solve a classroom problem created by plays with little performance history or connection to familiar theatrical styles. It will also test the limitations of the TEI scholarly encoding guidelines by exploring how, in the case of play scripts, building hypermedia resources requires creative, user-oriented strategies of encoding. The research-creation program will illustrate how contemporary artists can engage with historical works, while shedding light onto the theatrical creative

process. Finally, our *Resonant Response* to Joanna Baillie's Drama will combine scholarly research on Romantic drama, practice-driven analysis, the creation of new work, and hypermedia expertise.

This particular research-creation program is singular and innovative in its combination of academic close reading, dramaturgical analysis, dramatic writing and theatrical performance, filmed theatre, and a resolutely pedagogical preoccupation with a full and thorough exploration of the possibilities of hypermedia edition.

In addition to creating a prototype hypermedia edition, the project seeks to find out:

- what value performance annotations can add to a teacher's work with students on a seldom-performed play;
- how the Text Encoding Initiative's (TEI) scholarly encoding guidelines can best be customized to design hypermedia play editions;
- how the process of collaboration among faculty and students in humanities and communications disciplines can enrich understanding of technology's interaction with interpretation.

Integration of Distributed Text Resources by Using Schema Matching Techniques

Eckart, Thomas
teckart@e-humanities.net
Natural Language Processing Group, Institute of Computer Science, University of Leipzig, Germany

Pansch, David
dpansch@eaqua.net
Natural Language Processing Group, Institute of Computer Science, University of Leipzig, Germany

Büchler, Marco
mbuechler@e-humanities.net
Natural Language Processing Group, Institute of Computer Science, University of Leipzig, Germany

1. Scattered Landscapes

The world of humanities has seen an enormous growth in available digital text resources in the last decade. This development was driven by many factors like advancements in relevant technologies like OCR, increasing competence in the field of digital encoding and publication, and the spreading of widely accepted encoding formats. It is by now widely understood (among both researchers and funders) that publications and created resources have to be standardized to ensure their relevance for future work and their (re-)use in a linked data environment.

More and more projects with partially very specific research questions are working on the encoding of their results in various (mostly XML based) formats. Encoding standards and formats were established that are widely used and supported by various tools and a helping community. In the field of encoding textual resources the standards of the Text Encoding Initiative (TEI) and their various dialects are well represented. To cover a wide range of data it is common to create a specific dialect that fits the own data best without losing all compatibility with other projects. As a consequence, various encoding variants exist that cope with similar data but create different schemata to represent them. A drawback of this specialism are problems with aggregating existing data stocks to one global resource: Even combining solely meta data of editions of the same work becomes an expensive (since labor intensive) task. Creating true

hypertextuality in digital libraries (Berti et al., 2009) that will massively connect resources in a distributed infrastructure will intensify this problem. Hence data integration will gain relevance in the field of distributed resources.

Since data storage solutions like relational database management systems are often used (for example when fast access to or complicated requests on the data is necessary) this issue does not only apply on the XML data model. These systems often use ETL-procedures (Extract, Transform, Load) to gain uniform and comparable data. The key problem remains the same: A lot of time is spent to gain a clean stock of data and consistent meta data. Experiences show that especially at projects using quantitative approaches, with a demand for large (and as homogeneous as possible) data sets, up to one third of all human resources are needed to overcome different kinds of heterogeneity.

This paper concentrates on the question how existing schema matching techniques that are established in data warehousing and information integration can be used to identify identical structures of different editions of the same source material. Therefore a high similarity of the content can be assumed, whereas structural and semantic heterogeneity prevents fast integration. As all modern storage solutions rely on schema definitions, it was not the task to identify corresponding element instances of two documents but to identify correspondences between collections of documents. For this reason in the following the term 'element' will be used for the set of all elements having the same position in one schema (for example all *TEI/teiHeader/fileDesc/titleStmt/title*elements in a set of XML files or all values for one column *work-name*in a relational database table).

This approach has two advantages: generic profiles for every schema element can be created (thus minimizing the effects of outliers) and computational time is reduced by minimizing the number of comparisons that have to be made to find useful schema mappings.

To illustrate the procedure different versions of the Duke Databank of Documentary Papyri (DDbDP) were used. These include the Perseus version of the DDbDP, its EpiDoc encoded equivalent (Epiduke) and an extraction from the latter, stored in a flat relational schema. These data are only to be seen as a first testing environment. Further evaluation on other text types is in progress.

2. General Approach

The whole process of finding corresponding elements or larger element structures can be separated into three major working steps:

1. Fingerprinting: By using various features a fingerprint is created for every element, taking different element properties into account.
2. Linking: Elements of both schemata are chosen pairwise that are likely matching candidates.
3. Scoring: Every linked pair is scored by a similarity measure.

To identify corresponding elements of two schemata, for every addressable element various features are used. These features address different types of similarity like structural similarity (with the focus on schema information) or semantic similarity (with the focus on elements' content). Most of the used features do not depend on specific structures or access methods, hence every addressable element can be used and compared. As a consequence XML or SGML documents, columns in a relational database or every other (semi-)structured input schema can be used. Existing works have shown that using only a single feature is not sufficient to identify similarity (Algergawy et al., 2009). For this reason all measures are combined and normalized by a weighted sum.

As there is a wide range of syntactic and semantic ambiguities it is unlikely to achieve a full automated matching. Hence it is the goal to establish an integration procedure that allows a more efficient handling of new data resources to minimize the effort of integrating these resources into an existing data stock.

3. Methodology

3.1. Used Features

A wide range of different features are known in the field of data integration. Some of these make use of structural schema information (schema based) while others use the elements' content (instance based). A constraint based approach checks the type and limitations of data, e.g. the domain of numbers or the differences in cardinality or uniqueness of elements. These features work on different levels: some concentrate on the combination of elements, their hierarchy or their number of child nodes (structure level), while others focus on individual elements and their attributes (element level).

The following features were tested on their usefulness for the described problem. Table 1 gives a short overview of used approaches and their classification (based on Rahm et al., 2009).

- *Name similarity*uses the Levenshtein distance to compute string similarity of database column names, respectively XML element names. For example an element name *authors*has a distance of 1 from an element named *author*, but a distance of 5 from the string *work*.
- *Path similarity*compares the structural depth of elements, under the assumption that similar elements have similar positions (and therefore similar distances to the root element) in their respective schema.
- *Cosine similarity*uses the Vector Space Model by representing the content (i.e. the occurring terms) of every element as a vector in a high dimensional vector space. To reflect different importance of terms (for example stop words versus domain-specific keywords) all terms are weighted by using the *tf.idf*measure (Salton et al., 1988). The result vectors are compared using the cosine similarity: $sim_{cos}(p_1, p_2) = \frac{v_{p1} \cdot v_{p2}}{|v_{p1}| \cdot |v_{p2}|}$.
- *Dice coefficient*calculates the ratio of words, that appear in both compared elements to all occurring words: $sim_{dice}(p_1,p_2) = \frac{2 |W_{p1} \cap W_{p2}|}{|W_{p1}| + |W_{p2}|}$. For example an element that contains the words {bank, money, account, credit} is similar to an element containing the words {bank, money, account, financial} (Dice coefficient = 0.75), but less similar to an element containing the words {bank, river, water} (Dice coefficient ~ 0.29)
- *Frequency similarity*uses the assumption that similar content is encoded by a similar number of elements. Therefore this measure produces a high value if the number of occurrences of the compared elements are similar.
- *Content type*compares the ratio of numbers to letters. Hence an element with mostly numbers becomes dissimilar to an element containing mostly textual data.

All results were normalized to the interval [0,1] (where necessary), `0' corresponding to no similarity and `1' to identity.

Features that address the element's content use all available data: For example the union of all text addressable with the same XPath expression in a collection of XML files or all data in a column of a relational schema.

Similarity measure	Schema-based	Instance-based	Constraint-based	Structure level	Element level
Name	x				x
Path	x			x	
Cosine		x			x
Dice		x			x
Frequency	x		x	x	x
Content type		x	x		x

Overview of used measures and their classification

3.2. Linking

Experiments have shown that many elements in the chosen XML collections occur very rarely. Therefore only elements were considered that occur in at least 50 percent of all documents of the respective collection to reduce the computation time. All other elements of both compared data sets were linked with each other.

In general a more sophisticated approach would be useful to minimize the number of comparisons. This holds especially true in a distributed environment where network response and transmission time is a limiting factor. This was not considered in this work as the focus was on identifying useful features and all resources were locally available.

3.3. Scoring and Results

The values of all similarity measures are combined by a weighted sum, yielding a similarity value between `0' (no similarity) and `1' (identity). Starting by identical weights for all measures the weights were iteratively adjusted to enhance the matching precision.

The results show that especially the instance-based approaches (Dice and Cosine) were successful for identifying matching elements. These measures worked well on both XML-XML and relational database-XML comparisons. All other measures turned out to be strongly dependent on the compared formats.

Especially structural differences between optimized (and redundancy-free) relational schemata and XML documents prevent good results when relying solely on schema information: in these cases matching precision drops significantly. Only instance-based measures (Dice, Cosine, and content type similarity) achieved good results, whereas all other measures could be ignored (hence weighted with 0).

A more balanced result shows in the XML-XML analysis: as both document collections are based on subsets of TEI formats many relevant elements are similar regarding their name and position in the XML structure tree. Therefore especially name similarity turned out to be a useful indicator for matching elements. Nonetheless again instance based measures performed best on these comparisons.

3.4. Visualization

A graphical user interface was created to visualize the pairwise similarity of elements. In a matrix (c.f. Figure 1) every row stands for an element of the document collection 1 and every column for an element of document collection 2 (or database columns in case of a relational schema). Each cell contains the weighted sum of similarity values for the two respective elements. High certainty values are emphasized with a strong green color. A tooltip on each cell gives additional information about the comparison results (used features and similarity values).

Figure 1 shows an excerpt of a result matrix where a relational database is compared to a collection of XML files. Every column represents a database column of the extracted version of the Epiduke, every row represents a path in Perseus' DDbDP XML files. As an example the element *placeName*from the collection of XML files has a strong similarity to the *geography*column in the database, even though different element names were used. By analyzing their content it becomes obvious that there is a strong extensional overlap between these elements' content. On the other hand there are no similar XML elements for the database column *female*(information about the author's gender) and *work_id*(a database-specific identifier for every work). This is due to the fact that both elements were created during the extraction process (or the following post-processing) and therefore have no corresponding elements in the original material.

	author	dating	female	geography	work_id	corpora_a...	sentence
TEI	0,15321	0,06600	0,00075	0,05813	0,00169	0,15276	0,42425
teiHeader	0,35318	0,05009	0,01813	0,05964	0,00702	0,35238	0,06563
fileDesc	0,34610	0,04906	0,01609	0,04641	0,00359	0,34491	0,05296
titleStmt	0,35476	0,05359	0,00556	0,05000	0,00000	0,89750	0,05554
title	0,35754	0,05637	0,01667	0,05000	0,00000	0,89750	0,06248
publicatio...	0,35562	0,06225	0,00706	0,04960	0,00706	0,35035	0,05323
authority	0,08333	0,05915	0,00000	0,05556	0,00556	0,06765	0,04998
idno	0,70821	0,04639	0,01441	0,03559	0,01441	0,65902	0,04192
availability	0,05123	0,05775	0,01127	0,05539	0,00711	0,05294	0,04733
text	0,05801	0,08135	0,00868	0,06951	0,00157	0,05262	0,59412
body	0,04968	0,07302	0,00035	0,07506	0,00871	0,05262	0,58162
head	0,04623	0,52029	0,02044	0,45533	0,01649	0,05211	0,05345
date	0,05636	0,97401	0,01864	0,05488	0,00197	0,05391	0,06101
placeName	0,05556	0,05488	0,01667	0,94999	0,00000	0,05588	0,06119

NameSim: 0.0, PathSim: 1.0, VectorSim: 0.9296503, DiceSim: 0.6497923, (Abs)FreqSim: 0.002165949, (Rel)FreqSim: 8.460738E-6, ContentTypeSim: 0.98647714

Figure 1: *Graphical output for comparison of epigraphic data hold in Perseus' DDbDP XML and an extracted version of the Epiduke.*

4. Conclusions and Further Work

Various comparisons have shown that especially semantic approaches are promising for identifying similar elements. Apparently these measures' results will degrade when data is compared with a very small semantic overlap (like editions of different domains or languages). As a consequence structural information could be taken into account. For the analyzed data these measures proved useful where complex structures exist, but failed on flat relational schemata.

The existing system is only to be seen as a prototype that will be extended in the future. The further focus will be set on adding new features and analysis of an extended set of input. It is expected that more efficient and domain-specific profiles can be created that will be basis for useful weights and combinations of features. Additionally it is expected that the results can be improved by using a more in-deep view on the data (like identification of key words) or further exploitation of structural information (by including schema definitions into the classification process).

References

Algergawy, A., Nayak, R., Saake, G. (2009). 'XML Schema Element Similarity Measures: A Schema Matching Context'. *Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009 on On the Move to Meaningful Internet Systems OTM.* Vilamoura, Portugal, 2009, pp. 1246-1253.

Berti, M., Romanello, M., Babeu, A., Crane, G. (2009). 'Collecting fragmentary authors in a digital library'. *9th ACM/IEEE-CS joint conference on Digital libraries.* Austin, TX, USA, 2009, pp. 259-262.

Bizer, C., Heath, T., Berners-Lee, T. (2009). 'Linked Data - The Story So Far'. *International Journal on Semantic Web and Information Systems.* 1-22.

Epiduke Online publication. `http://epiduke.cch.kcl.ac.uk` (accessed 1 November 2010).

LaQuAT: Linking and Querying Ancient Texts. `http://www.kcl.ac.uk/iss/cerch/projects/completed/laquat` (accessed 28 November 2010).

Pansch, D. (2010). *Data Integration Methods for Structural Heterogeneous Data in an eHumanity Context.* Leipzig.

Rahm, E., Bernstein, P.A. (2001). 'A survey of approaches to automatic schema matching'. *VLDB Journal.* 10.

Salton, G., Buckley, C. (1988). 'Term-weighting approaches in automatic text retrieval'. *Information Processing and Management.* 5: 513--523.

TEI: Text Encoding Initiative. `http://www.tei-c.org/index.xml` (accessed 23 November 2010).

Do Birds of a Feather Really Flock Together, or How to Choose Test Samples for Authorship Attribution

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

Rybicki, Jan
jkrybicki@gmail.com
Pedagogical University of Kraków, Poland

In the house of non-traditional authorship attribution are many mansions, or methods, based on statistical analysis of authorial style. They all compare text samples of disputed or unknown authorship to texts written by known authors, or "candidates". The degree of similarity or dissimilarity between samples allows informed guesses on the possible authorship of a given text. The so-called machine-learning methods are supposed to be among the most effective; they include Support Vector Machines, Nearest Shrunken Centroid classification, Burrows' Delta and so on (for a comparison of their effectiveness cf. Jockers and Witten 2010).

The general feature of the methods in question is a two-step supervised analysis. In the first step, the traceable differences between samples constitute a set of rules, or a classifier, for discriminating authorial "uniqueness" in style. The second step is of predictive nature – using the trained classifier, one can assign other text samples to the authorial classes established by the classifier; any disputed or anonymous sample will be assigned to one of the classes as well.

The procedure described above relies on a pre-processed corpus of samples. Namely, the clue is to divide all the available text samples into two groups: primary (training) set and secondary (test) set. The first set, being a collection of texts written by known authors ("candidates"), serves as a sub-corpus for finding the best classifier, while the second set is a pool of texts of known authors, anonymous samples, disputed ones and so on. The better the classifier, the more samples from the test set are attributed correctly and the more reliable the attribution of the disputed samples.

Such procedures have been successful in social and medical studies; no wonder, then, that it soon made

its way into authorship attribution. Yet, contrary to the former applications where the researcher usually enjoys a high number of test samples (e.g. patients), authorship attribution has to struggle with a limited number of samples available to train a convincing classifier. This makes the classifier sensitive to statistical error. What is more, the generally-accepted division of data studied into a training set and a test set further limits the texts that can be attributed.

This sensitivity of machine-learning classifiers to the choice of samples in the training set has already been observed (Jockers and Witten 2010: 220). Intuition suggests composing the training set from the most typical texts (whatever "typical" means) by the authors studied (thus, for Goethe, *Werther* rather than *Farbenlehre*). In practice, this can be quite complicated: in a small corpus, to change a single training set sample for another can upset the delicate mesh of interrelationships between all other texts. This potentially heavy impact on the effectiveness of attribution tests has not been lost on Hoover: "As a reminder of how much depends upon the initial choice of primary and secondary texts, consider what happens if the same 59 texts are analyzed again, but with different choices for primary and secondary texts [...]. If the analyses that are the most successful with the initial set are repeated, Delta successfully attributes only 16 of the 25 texts by members of the primary set" (Hoover 2004a: 461).

Last but not least, any manual selection of texts to both sets must be highly arbitrary. To further quote Hoover: "The primary novels for this test are intentionally chosen so as to produce poor results, but one might often be faced with an analysis in which there is no known basis upon which to choose the primary and secondary texts, and nothing prevents an unfortunate set of texts like this from occurring by chance" (Hoover 2004a: 461-62).

Machine-learning methods routinely try to estimate the potential error due to incorrect choice of the training set samples. This cross-validation consist in a few random changes to the composition of both sets, followed by a comparison of the classifier's success, ten-folded cross-validation being the standard solution (Tibshirani *et al.* 2003: 107; Baayen 2008: 162; Jockers and Witten 2009: 219). The question arises whether ten trials are sufficient for a classifier which, based on but a few samples, can be unstable.

Assuming that the training set contains 10 samples by 10 authors, and the test set another 10 samples by these authors, there are $2^{10} = 1024$ possible combinations of members of the training set. For a corpus of 60 novels by 20 authors, this number becomes so large that testing all possible permutations of both sets is unrealistic. Instead, the impact of the composition of the training set on attribution success can be assessed basing on several hundred random permutations; this can be done with a variety of bootstrap procedures (Good 2006).

To test this problem, we have selected several corpora of similar size and similar number of authors studied (with the obvious caveat that any comparison between different languages can never be fully objective). For each of these corpora, we have performed 500 controlled attribution experiments, each with a random selection of the training and the test sets. We have compared the number of correct authors guessed, with the hypothesis that the more resistant a corpus is to changes in the choice of the two sets, the more stable the results.

All tests featured the simplest, the most intuitive and the most frequently used of machine-learning attribution methods: Burrows's Delta (Burrows 2002; Hoover 2004b). Delta was run for 100 MFWs, then for 200 and then, at increments of 100, all the way to 2000 MFWs. This was performed at five different culling settings (0-100% incrementing by 20), giving a total of 1000 results, and a mean of these was recorded. The above procedure was then repeated for 500 random permutations of the texts in the training set. The density function was estimated for the final results thus obtained.

It can be assumed that the distribution of these 500 final results should be Gaussian rather than anything else. The peak of the curve would indicate the real effectiveness of the method, while its tails – the impact of random factors. A thin and tall peak would thus imply stable results, i.e. those resistant to changes in the primary set.

The analysis of the results begins with the corpus of 63 English novels by 17 authors. As expected, the density of the 500 bootstrap results follows a (skewed) bell curve (Figure 1). At the same time, its gentler left slope suggests that, depending on the choice of the training set, the percentage of correct attributions can vary, with bad luck, to below 90%.

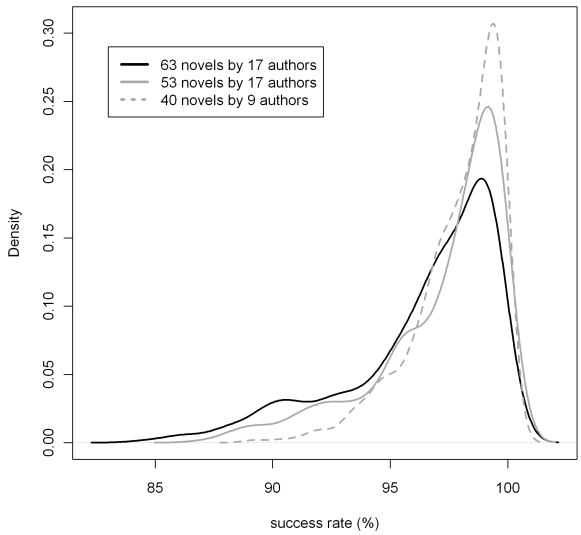


Figure 1. Density (vertical axis) of attributive success percentage rates (horizontal axis) in the English novel corpus

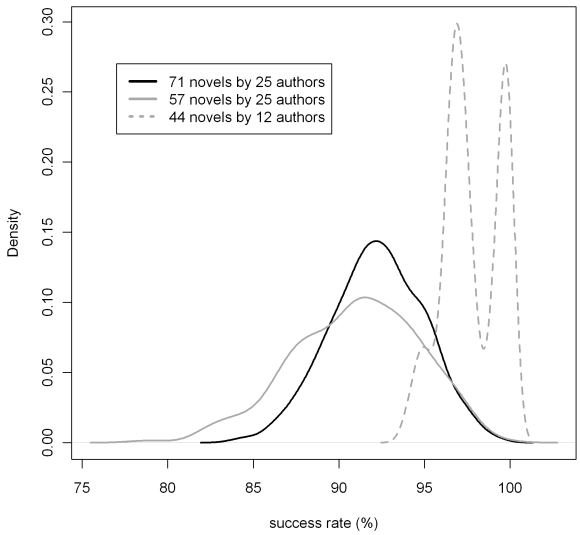


Figure 2. Density of attributive success rates in the French novel corpus

It is quite natural that the stability of the results might also depend on the number of authors and/or texts analyzed. The same Figure shows that, with fewer authors, a higher number of texts has no significant impact on the stability of the results at any permutation of both sets (the dashed line), as already observed by Hoover and Hess (2009: 474). With more authors (i.e. when guessing becomes more difficult), the curve widens and a perfect match is even less frequent.

And this is still good accuracy and a fairly predictable model. However, it has to be remembered that Delta has been shown to be somewhat less perfect in other languages (Rybicki and Eder 2011).

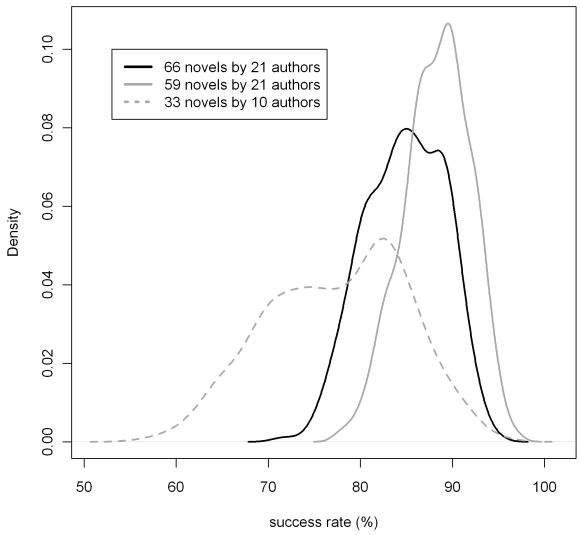


Figure 3. Density of attributive success rates in the German novel corpus

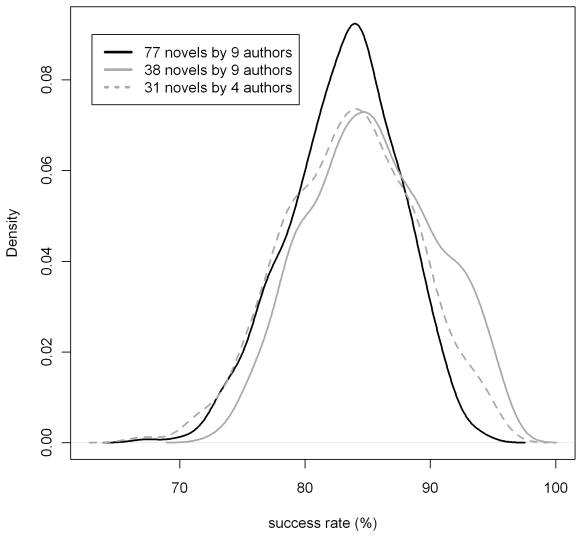


Figure 4. Density of attributive success rates in the Italian novel corpus

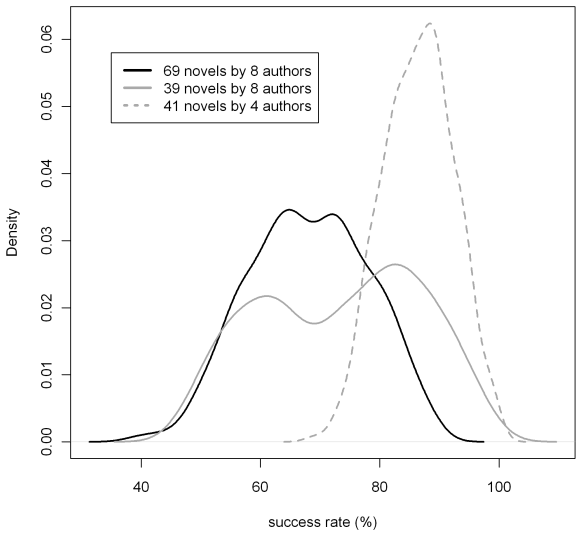


Figure 5. Density of attributive success rates in the Polish novel corpus

Indeed, the discrepancies in Figures 2-5 seem to question the validity of attribution tests based on arbitrary choice of training sets. Although peaks for some combinations of numbers of texts and authors may be at acceptable levels, the left slopes of the curves tend towards dangerously low values; and the wide tails of the curves show that a high success rate outliers might be a stroke of luck rather than a consequence of the method, the data and the statistical assumptions – the most ominous memento appearing here from the inexplicable dispersion in the corpus of 39 Polish novels by 8 authors (Figure 5, grey solid line). Therefore, the ideal authorship attribution situation is not only that of many texts by many authors; it is equally important to assess the validity of the training set with a very high number of trials. This seems to be the only way to escape the quandary of arbitrarily naming each author's "typical" text.

References

Burrows, J. F. (2002). ''Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship'. *Literary and Linguistic Computing.* 17(3): 267-87.

Good, P. (2006). *Resampling Methods.* Boston, Basel, Berlin: Birkhäuser.

Hoover, D. L. (2004a). 'Testing Burrows's Delta'. *Literary and Linguistic Computing.* 19(4): 453-75.

Hoover, D. L. (2004b). 'Delta Prime?'. *Literary and Linguistic Computing.* 19(4): 477-95.

Hoover, D. L., Hess, S. (2009). 'An Exercise in Non-ideal Authorship Attribution: The Mysterious Maria Ward'. *Literary and Linguistic Computing.* 24(4): 467-89.

Jockers, M. L., Witten, D. M., Criddle, C. S. (2008). 'Reassessing Authorship in the 'Book of Mormon' Using Delta and Nearest Shrunken Centroid Classification'. *Literary and Linguistic Computing.* 23(4): 465-91.

Jockers, M. L., Witten, D. M. (2010). 'A Comparative Study of Machine Learning Methods for Authorship Attribution'. *Literary and Linguistic Computing.* 25(2): 215-23.

Rybicki, J., Eder, M. (2011). 'Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?'. *Literary and Linguistic Computing.* 26: (forthcoming).

Tibshirani, R., Hastie, T., Narashimhan, B., Chu, G. (2003). 'Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays'. *Statistical Science.* 18: 104-17.

Knowledge and Reasoning: Connecting Scientific Data and Cultural Heritage

France, Fenella G.
frfr@loc.gov
Library of Congress, United States of America

Toth, Michael B.
mbt.rbtoth@gmail.com
R.B. Toth Associates

Uncovering and detecting patterns of information in library and museum collection items requires the integration of scholarly and scientific data analyses. Research into the materials substrate (paper, parchment, etc) and media (ink, pigment, colorant) that comprise the historic object inform the scholar as to the provenance of a particular document. This may be through characterization and identification of the media and substrate on which the information is contained, or watermark identification that can effectively date the document to a specific time period. The challenge of linking disparate materials characterization and identification databases and scientific analyses is a critical research issue requiring the development of a knowledge representation (KR) to facilitate interpretation that enables this humanities research. With cultural heritage objects, the KR must maintain linkage between the original document that contains a wealth of knowledge stored contextually, and digital surrogates that represent the document. These links will be explored in reference to cultural treasures of South America, where the integration of scientific analyses and historic scholarly information led to the generation of knowledge that enriched the contextual interpretation of the original object. Scientific analyses of textile treasures from Llullaillaco provided an understanding of their use and purpose, while environmental information of Pre-classic Mayan structures in the Mirador Basin allowed an assessment of preservation requirements. Investigations of maps with specific links to South America yielded information on the source of pigments and geographical location of their original creation.

Historic documents and cultural heritage objects do not generally lend themselves to ease of context analysis, since documentation about the creation of the document is not readily available. Discovery of data with more detail about the context and circumstances that surround the creation of the artifact allows researchers to visualize information previously not detected. This visualization is achieved in Library of Congress studies through advanced spectral imaging techniques that incorporates data from both visible and non-visible regions of the spectrum to create an integrated “digital cultural object” (DCO). The additional contextual information is not apparent in conventional digitization techniques for these objects, so the integration of the spectral data assists in mining the layers of data stored within the objects. In this way the DCO provides a range of information that allows a shift from the use of interpretive virtual heritage applications that focus on the artistic rather than the investigative and inferential, towards the development of interdisciplinary scientific data analyses as part of cultural heritage humanities scholarly studies. In this way, cultural heritage, science and technology are intertwined, advancing the capacity to mine and analyze historic data from multiple viewpoints. Interaction and interpretation of these additional dimensions allow the description of new relationships between constituent elements, connecting patterns and mining the data for trends, and correlating formerly disparate components. For example, the new identification of pigments in an object that come from a different geographical location, can then suggests trade routes and exchange of materials and artistic techniques. The *UNESCO Charter on the Preservation of Digital Heritage* has recognized the importance of digital versions of cultural materials, referring to digital heritage as “resources of information and creative expression” being “increasingly produced, distributed, accessed and maintained in digital form.”

The composite of images of the maps and textiles that forms the new DCO is related to, but distinct from the originals. This digital object is not a surrogate for the original, but provides new knowledge through the integration of scientific analysis of these cultural heritage objects within libraries, museums and archives. The range of data this new digital object contains enhances interaction between a range of professions, allowing multidisciplinary collaboration for integration of preservation, sociological and cultural information. Digitally generated and accessed data for these maps and textiles balances the opposing goals of libraries and museums and optimizes preservation of the original objects, while increasing access to information from the original. Hyperspectral imaging allows the DCO to create a new interpretation of the original object, as apparent from the 3-dimensional reconstruction of the original woodcut of the Waldseemüller 1507 world. Since these original manufacturing materials no longer exist,

the DCO allows the representation of this new scientific knowledge to link with geography and map curatorial knowledge of the era, and assist a greater understanding of the printing techniques and possible location of where these materials originated. This was the first time America was referred to as America, on any map, printed or manuscript, with a unique perspective of South America.

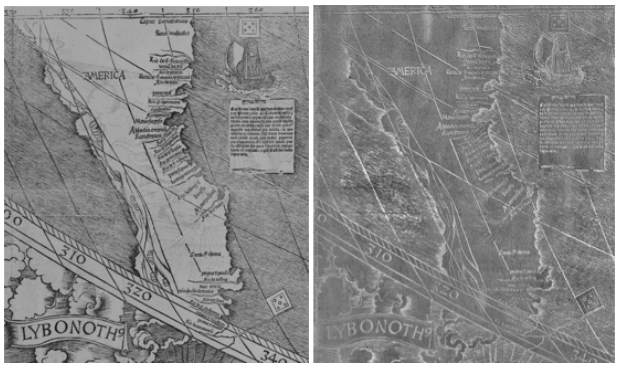


Figure 1. LHS: Section of Printed Sheet of South America, RHS: 3D Rendering of Woodcut

Image capture and processing of these and other objects is important for the interpretation of cultural heritage by allowing layers of data to be analyzed and linked. This offers an archeological examination of the information strata, the materials, inter-connections between the materials used, discovered text and information, and the relative associations between the components of the artifact. The creation of an image data-cube deconstructs layers of data into discrete components, while conversely also integrating and utilizing the application of scientific methods to the recognition of areas of interest within the artifact. The explanations generated from these processes expand the associations and extracted knowledge of the original cultural heritage objects.

There is dynamic interaction between re-examination of the original or source materials and the DCO, with raw and processed data that can enhance obscured and specific features. Inter-connections and relationships between the source and the generated interpretations based upon analysis of generated data are an iterative process. The process of interpretation relies upon the use of implicit assumptions, inferences or internal filtering. For both scientific and scholarly researchers, these processes are often based upon prior knowledge and experience. Additional filtering for scientific analyses are introduced by reference databases that match known reference materials with “pixel” samples taken from the spectral images to non-invasively characterize materials. These categorizations are objective in nature and reduce the potential error sometimes introduced by subjective assumptions. However, the essential element that should not be overlooked in this iterative process is the power of strong collaborations between a range of disciplines. In 1999 on Llullaillaco's summit, an Argentine-Peruvian expedition co-directed by Johan Reinhard and Argentine archaeologist Constanza Ceruti found the perfectly preserved bodies of three Inca children, sacrificed approximately 500 years earlier. These were accompanied by a range of textiles, figurines and other ceremonial sacrificial materials. A close collaboration between the American Museum of Natural History, preservation scientists, engineers from Argentina and the USA, conservators and curators was critical to ensuring the preservation of these unique materials. The interweaving of the assessment of current condition of the cultural artifacts such as strength and chemical degradation to aid their preservation, with scholarly analysis of the unique patterns of construction enabled the knowledge of both to be linked so as to create a rare collaboration of the forensic type recovery of the history of these materials. This is the highest Inca burial so far discovered and the world's highest archaeological sit. There was intense pressure for exhibition of the materials so close collaboration between all parties to gain as much scientific and cultural information about the origins of the materials and the mummies was a unique component in ensuring their longevity while on exhibit. This sharing of data aided the control of exhibit conditions in South America and long-term management of the materials to allow further studies.

The critical component in the generation of knowledge in these studies is recognizing the importance of skilled people and work processes that efficiently add value and meaning. In order to analyze the original source material, high resolution spectral imaging creates the DCO, and data processing and further scientific analyses revolve around the interactions of the people involved – preservation scientist, curator, scholar and technology specialist. This iterative process is reliant not only upon the effective use of technology to assimilate process and disseminate the information with standardized processing, metadata and data management, but also the quality of the collaborative interaction between professionals from different fields.

While the relationship between the original and the DCO is provided through metadata that maintains the spatial links between the scientific data, the original material, and the new knowledge generated from these linkages, the integration of diverse opinions and perspectives is an integral component of this new knowledge generation. Standardization of file

formats and structures across these different fields provides a method of ensuring continued access and integration to the information, by maintaining and creating effective associations while generating new knowledge. The above examples illustrate the requirement for this standardization of both scientific and scholarly files, to enable true international collaboration and sharing of resources between countries and disciplines. These protocols can support effective data exchange with conventions that provide a local structure for a scientific data network. This sustains diversity in scientific research and scholarly studies. This requirement for a structural framework for cultural heritage institutions allows both user access and functional usability of information to support research.

Effective visualization of these data connections is essential for further associations, with access to both spatial and temporal data for the maps, textiles and other objects directly linked back to the original source material through the DCO. Visualization tools and interfaces offer potential for open dialogue between multidisciplinary fields and ease of navigation through layers of data, since knowledge generation is reliant on the cohesive interaction and collaboration between science and humanities researchers. The knowledge representation and underlying interpretation system needs to be appropriate for the application and the types of analyses that are needed for integrating and expanding cultural heritage research. The development of XML-based knowledge representation languages and standards include the Resource Description Framework (RDF). A major benefit of RDF is the ability to utilize the features that facilitate the merging of data, even if the underlying schemas differ. The further advantage for this heritage science application is the capacity to support the evolution of schemas over time, allowing both structured and semi-structured data to be combined and shared across different areas within the application. The large volumes of digital data generated require a repository that can cope with a diverse range of object types, as is true for with collections of data used for scientific analyses. Data files from these objects include images, spectra, reports, and other extant files, requiring additional metadata mechanisms to accommodate a range of data, while retaining access and associations. An RDF model offers greater flexibility and opportunity for integration across various disciplines, since it is designed to accommodate multiple ontologies with a structured approach. The model or ontology being exploited within this structure allows us to coordinate a wide range of data and file types with extensive metadata and multiple data formats. For example, metadata from the scientific instrument is included so the measurement could be reproduced, and in addition the data can be imported in multiple formats – as spectra for visual representation and as a .csv standardized sharable data file.

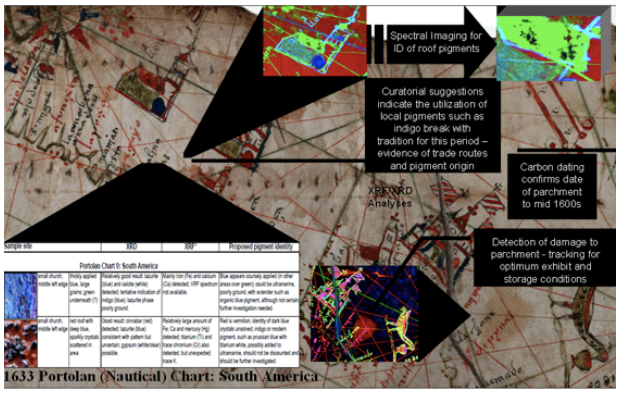


Figure 2. Visual Representation of Layers of Digital Object Data

Scriptospatial mapping of the textiles and maps involves an accurate coordinate system that links scientific and scholarly analyses to the DCO, and allows inferences to be drawn to generate new knowledge. This approach to viewing the DCO in relation to multiple dimensions applies an essentially archeological methodology toward uncovering and interconnecting information strata of cultural heritage artifacts. Utilizing an object-oriented approach in conjunction with the data layer allows the mapping of spatial and temporal data with increasing complexity. Examining and explaining the physical, spectral and chemical properties of the maps and textiles permit the humanities scholar to link these scientific analyses to the social aspects of how they were created. These links therefore create meaningful scientific outcomes of the content: When obscured or faded text can be retrieved; inks and pigments characterized and traced to specific geographical locations; analysis of the intensity of handwriting imparts understanding of the author's original intent; and the provenance and source of paper is gleaned through the capture and analysis of the watermark in the paper.

A continued focus on collaboration between people, data and processes is a major factor in promoting access and integration of scientific and humanities research, emphasizing the importance of linking the original artifact with digital tools and techniques for visualizing and disseminating new knowledge in the arts and humanities. Concentrating on generating new knowledge from content derived from these maps, textiles and other objects enhances the importance of the DCO. This allows improved access, interpretation and preservation of fragile items of significant cultural

heritage. However the extracted information is only as important as the strength of the collaborative partnerships set up to create a constant iterative loop for access to and interpretation of new scholarly and scientific information. This requires a strong and committed association between previously disparate fields to incorporate and share the generation of new knowledge by mining additional data and forging new advances in humanities research. These related but previously disparate disciplines comprise humanities scholars, scientists, researchers, technology and data management specialists to form an open yet interconnected digital exchange of humanities research.

References

Cameron, F, and Kenderdine, S. (eds) (2007). *Theorizing Digital Cultural Heritage: A critical discourse.* Massachusetts, USA: Massachusetts Institute of Technology Press.

Emery, D, France, F. G., and Toth, M. B. (2009). 'Management of Spectral Imaging Archives for Scientific Preservation Studies'. *Archiving 2009, Society for Imaging Science and Technology.* Arlington, VA, May 4-7, pp. 137-141.

Esteva, M., Trelogan, J., Rabinowitz, A., Walling, D. and Pipkin, S. (2010). 'From the Site to Long-term Preservation: A Reflexive system to Manage and Archive Digital Archeological Data'. *Archiving 2010, Society for Imaging Science and Technology.* The Hague, June 1-4.

France, F.G. (2010). 'Spectral Imaging and Non-Invasive Characterization of Manuscripts'. *Eikonopoiia: Symposium on Digital Imaging of Ancient Textual Heritage.* Helsinki, Finland, October 28-29, 2010, pp. 51-64.

France, F.G., Christens-Barry, W., Toth, M.B., Boydston, K. (2010). 'Advanced Image Analysis for the Preservation of Cultural Heritage'. *22nd Annual IS&T/ SPIE Symposium on Electronic Imaging.* San Jose, January 2010.

France, F. G., Emery, D., and Toth, M.B. (2010). 'The Convergence of Information Technology, Data and Management in a Library Imaging Program'. *Library Quarterly special edition: Digital Convergence: Libraries, Archives, and Museums in the Information Age*.

France, F.G., Roussakis, V., Lissa, P., Xamena, M, Santillán, P, Capero de Larrán, M., Doña, G and Ammirati, G. (2005). 'Textile Treasures of Llullaillaco'. *North American Textile Conservation Conference.* Mexico City, November 2005, pp. 25-30.

Museum of High Altitude Archaeology (MAAM), Salta. `http://maam.culturasalta.gov.ar/` .

Schreibman, S. and Hanlon, A.M. (2010). 'Determining Value for Digital Humanities Tools: Report on a Survey of Tool Developers'. *Digital Humanities Quarterly.* 4 (2).

Svensson, P. (2010). 'The Landscape of Digital Humanities'. *Digital Humanities Quarterly.* 4 (1).

UNESCO Charter on the Preservation of Digital Heritage. `http://unesdoc.unesco.org/` .

Approaching the Coasts of *Utopia*: Visualization Strategies for Mapping Early Modern Paratexts

Galey, Alan
alan.galey@utoronto.ca
University of Toronto, Canada

1. Introduction

No one reads or writes a book alone. Proof may be found in the paratextual letters and other prefatory material that often accompanies a book into the world to meet its readers. This is especially true of early modern books, whose prefatory letters stand as a threshold where the book's material and symbolic production come together—sometimes as a well-executed plan (ex.: the 1518 editions of More's *Utopia*), and sometimes as a collision of intentions (ex.: the 1590 edition of Spenser's *Faerie Queene*, or the 1623 Shakespeare First Folio). However, print-based methods of representing paratextual networks—especially in their temporal dimension, across multiple editions and translations—tend to emphasize the published book as product, at the expense of the book as a process. This paper takes the textual tradition of Thomas More's *Utopia*, with its unfolding process of paratextual change between editions, as a test case for the design of an open-source interface component to help digital editors visualize networks of paratexts in early modern books.

The study of paratexts has been reinvigorated in recent years, crossing national and period boundaries in the tradition of Gerard Genette's Paratexts, but more recently drawing energy from intersections between book history and digital humanities as interdisciplinary fields. Building on ongoing research on the digital modelling and visualization of paratexts and similar materials (Fekete & Dufournaud, 2000; Monella, 2008; Johnson), this paper argues that creating a digital visualization component for mapping has two benefits: first that a well-designed digital visualization can represent the structured fluidity and temporality of publication as a process that unfolds in time; and second, that the process of creating such a visualization affords a reciprocal opportunity to interrogate the digital tools and systems we use to represent the past.

This paper develops its argument in four sections:

1. Research context: archive and interface in digital textual studies
2. *Utopia* as a modelling challenge
3. Visualization strategies
4. Conclusion: visualization prototypes as essays

The small-scale project outlined in this paper is part of a larger project titled Archive and Interface in Digital Textual Studies: From Cultural History to Critical Design, funded by the Social Sciences and Humanities Research Council of Canada. This project is premised on two linked arguments. The first is that we need to understand how the figure of the archive operates in the cultural imagination, and how perceptions of digital archives are partly coded in advance by historical fears and desires about the continuity of knowledge. The second premise is that we need to develop traditions of digital interface design native to the humanities, and which reflect the humanities' uniquely valuable understanding of the cultural histories and material complexities of texts (Kirschenbaum, 2004; Drucker, 2011). The Archive and Interface project therefore seeks to bridge between textual studies and the design of digital interface tools in the humanities. It does so first by investigating the cultural history of the humanities archive through case studies such as *Utopia*, and second by building an online library of interface components designed to be part of that cultural history.

The interface library will focus on critical design strategies in four key areas: textual variation (when sources diverge in significant ways); paratexts (documents such as prefatory letters, often published with literary works in complex configurations); materiality (the relation of physical documents to digital versions); and performance (the relation of written texts to reading or enactment in physical spaces). This project's interface library focuses on putting humanities-designed interface components in the hands of electronic editors, and disseminating the methods by which those components may be created and modified by the larger community of computing humanists (the open-source model).

Granting that large-scale editing systems like Anastasia and Edition Production & Presentation Technology have their place in the digital humanities, this paper will argue that small-scale and (relatively) rapidly prototyped interface components, built by individuals or small groups with inexpensive tools, can reflect the critical, experimental nature of humanities design in ways that large projects cannot. Such

experimental capacity and structural flexibility is necessary in digital humanities projects if they are to learn from challenging materials like *Utopia*, and not simply take their representability for granted.

3. *Utopia* as a Modelling Challenge

The specific nature of *Utopia*'s challenge to a digital editor is that *Utopia*'s publication, as a collaborative project between early modern humanists, thematizes the very ideals and anxieties about the dissemination of knowledge that digital humanists have inherited. The book form—and by extension, the emergent network of humanist print culture—is not merely a delivery system for *Utopia*, but also one of its chief objects of scrutiny. In particular, *Utopia* simultaneously embodies and critiques the early modern archive with unusual perspicacity. This paper's analysis follows Warren Wooden and John Wall by approaching *Utopia* "not as an object of knowledge but as an occasion for an act of perception, an instrument for 'seeing' designed to call attention to what is involved in perception" (1985, p. 233). In this light, *Utopia* itself serves as a kind of visualization of the early modern humanist archive of texts.

We know from correspondence that More began writing *Utopia* with what is now the second book. From a reader's perspective, the text was written in reverse, with book 1 (written second) placed before book 2 (written first), and various prefatory materials (written last) accumulating before book 1. These prefatory materials—the letters, verses, diagrams, and maps that constitute the paratext of *Utopia*—increase and vary across the four editions published from 1516 to 1518, and change even more radically in subsequent early modern and modern editions.

It is common for modern editions to completely remove or reconfigure *Utopia*'s carefully constructed paratexts (Allen, 1963). Yet, paradoxically, there may be no single ideal configuration of paratexts, making the interpretation of *Utopia* as a material text especially reliant on representational methods and tools (Jardine, 1993; Leslie, 1998; Vallée, 2004; Kinney, 2005). Those tools have tended to take printed form, culminating in Terence Cave's printed guide to *Utopia*'s paratextual tradition. However, *Utopia* can be taken as an early experiment itself in humanist print culture, no less than the digital experiments we discuss at digital humanities conferences, which makes *Utopia* anything but passive material for representation and editing.

4. Visualization Strategies

Given *Utopia*'s playful, experimental nature, this paper argues for the need for a visualization strategy based not on static representation, as in traditional forms of data visualization that represents the results of analysis, but based rather on the idea of modelling. Unlike a static representation, a digital model embodies the process-friendly dynamism we expect of digital visualizations, but also a certain "rough-and-ready" form and heuristic flexibility (McCarty, 2004). These latter qualities we associate not with commercial software but with the community-designed code libraries found at SourceForge and similar places, which serve as the dissemination model for the Archive and Interface project's visualization components. (On humanities approaches to visualization, see Drucker, 2010, and other articles in the same special issue of *Poetess Archive Journal* on "Visualizing the Archive.")

The design methodology for the interface library will be consistent with Ajax web applications, a type of architecture that distributes processing between a server and a user's web browser, and which integrates well with XML databases and object-oriented design. HTML 5's new capabilities on the client side permit animation and time-based interactivity to be incorporated into interfaces in ways that used to be exclusive to Flash. This paper will include a brief demonstration of a browser-based interface component that uses animation to model paratextual change over time. The prototype presented here will rely on the encoding structures proposed by Monella (2008), but will approach the topic from the browser and interface side instead of drawing conclusions about the encoding.

5. Conclusion: Visualization Prototypes as Essays

The two strategies proposed above, modelling and browser-based design, will serve to illustrate the paper's broader conclusion that a digital humanities project organized around many small interface prototypes may yield more publishable components, respond more quickly to critical discourse about the material, and involve less risk than a single, large interface project. The nature of paratexts calls for an interpretive approach to digital representation, especially with material as complex as More's *Utopia*. This paper concludes that the humanistic critical tradition, embodied by *Utopia* as a collaborative project in publishing technologies of its own time, calls digital humanists to think of their visualizations and interface prototypes not just as finished tools, which emphasize

utility, but also as essays that put forward arguments and serve as pretexts for debate—like *Utopia* itself.

References

Allen, P. R. (1963). *Utopia and European humanism: The function of the prefatory letters and verses, Studies in the Renaissance, 10: 91–107.*. : .

(31 October 2010). *Anastasia.* . http://anastasia.sourceforge.net/.

Cave, T. (2008). *Thomas More's Utopia in Early Modern Europe: Paratexts and Contexts.* . Manchester: Manchester University Press.

Drucker, J. (2011). 'Humanities approaches to interface theory'. *Culture Machine.* V. 12, pp. 1–20.

Drucker, J. (2010). 'Graphesis'. *Poetess Archive Journal.* 1–50.

Edition Production & Presentation Technology.. http://www.eppt.org/eppt/.

Fekete, J.-D., and N. Dufournaud. (2000). *Compus: Visualization and analysis of structured documents for understanding social life in the 16th century. Proceedings of the Fifth ACM Conference on Digital Libraries.* .

Genette, G. (1997). *Paratexts: Thresholds of Interpretation, trans. J.E. Lewin..* Cambridge: Cambridge University Press.

Jardine, L. (1993). *Erasmus, Man of Letters: The Construction of Charisma in Print.* Princeton, NJ: Princeton University Press.

Johnson, J.P.. *Pale Tour.* http://https://pantherfile.uwm.edu/johnso73/www/paletour/.

Kinney, A.F. (2005). *Utopia's first readers.. In Challenging Humanism: Essays in Honor of Dominic Baker-Smith.* T. Honselaars and A.F. Kinney (ed.). Newark, NJ: University of Delaware Press.

Kirschenbaum, M.G. (2004). *"So the colors cover the wires": Interface, aesthetics, and usability.. A Companion to Digital Humanities.* S. Schreibman, R.G. Siemens, and J. Unsworth. (ed.). Oxford: Blackwell.

Leslie, M. (1998). *Renaissance Utopias and the Problem of History..* Ithaca, NY: Cornell University Press.

McCarty, W. (2004). *Modeling: A study in words and meanings.. A Companion to Digital Humanities.* S. Schreibman, R.G. Siemens, and J. Unsworth (ed.). Oxford: Blackwell.

Monella, P. (2008). *Towards a digital model to edit the different paratextuality levels within a Textual Tradition. Digital Medievalist.* . http://www.digitalmedievalist.org/journal/4/monella/.

Vallée, J.-F. (2004). *Printed voices and written friendships in More's Utopia. Printed Voices: The Renaissance Culture of Dialogue.* D. Heitsch and J.-F. Vallée (ed.). Toronto: University of Toronto Press.

Wooden, W.W., and J.N. Wall, Jr. (1985). *Thomas More and the painter's eye: Visual perspective and artistic purpose in More's Utopia. Journal of Medieval and Renaissance Studies.* .

Is There Anybody out There? Discovering New DH Practitioners in other Countries

Galina, Isabel
igalina@unam.mx
Instituto de Investigaciones Bibliograficas, Universidad Nacional Autnoma de Mexico

Priani, Ernesto
epriani@gmail.com
Facultad de Filosofía y Letras, Universidad Nacional Autnoma de Mexico

1. Introduction

Digital Humanities (DH) has been described as an 'emerging' field for some time now but many (Borgman 2009; Friedlander 2009; Presner 2009) agree that this is a crucial moment for the discipline. The DH community now has established research methods, scholarly conferences and journals (Borgman 2009), DH centres and labs (Svensson 2010; van den Heuvel et al. 2010) and MA studies (Clement 2010). However, it could be argued that an important area for further development is the true internalization of the DH community. Up to now much of the discussion has centered on DH projects in a handful of mainly English speaking countries (Terras 2006). In order to consolidate as a discipline, an important challenge for the DH community is to extend its international reach and incorporate work from a broader range of academic institutions and languages.

One of the main issues of course, is attempting to integrate with groups of scholars that have not necessarily identified themselves yet as a community or do not even know that DH exists. Our project has four main objectives:

1. raise awareness about DH
2. identify key scholars and projects
3. investigate key local issues in the development of DH projects
4. consolidate a community and find ways of linking with the international DH community

The aim of this paper is to present our initial experience with two workshops as a methodological approach to investigating DH work in an unknown landscape and present a preliminary report on the DH situation in Mexico in particular.

To our knowledge this type of work has not been done before, in particular for Latin America. The fact that there has been little participation of Latin American scholars does not necessarily imply that no DH work is being done in this region but could be rather a lack of connection with the DH community. Japan for example, reports that despite having a long tradition in DH projects, one problem has been cooperating with similar projects overseas (Muller 2010).

2. Methodology

It is well documented that finding digital humanities resources (Dunning 2006; Pappa et al. 2006, Warwick et al 2006) and DH tools (Juola 2008) can be a difficult task and charting an unknown territory poses further challenges. We decided that workshops would allow us to both raise awareness of DH as well as serving as exploratory method to identify key scholars, projects and local issues. Additionally time and financial constraints were a major factor. At a later date we hope to use more quantitative methods such as a national survey.

Two workshops were carried out at the Universidad Nacional Autonoma de Mexico (UNAM). As there is no record of DH activity in Mexico it was difficult to define where to start and so we relied heavily on personal experience and contacts to produce a list of possible participants. We contacted everybody we knew who had experience working on a DH project and asked them in turn to invite other participants in order to generate a snowball effect. The invitation included a very broad definition of DH projects. The workshops generated a large amount of interest and were well attended.

In order to provide a framework for the discussion we identified seven key topics from the literature:

1. organizational context including institutional recognition and support (Siemens et al. 2010; Warwick 2008b)
2. planning and development intellectual
3. property and copyright (Rehm 2007); human resources and training (Warwick 2009)
4. dissemination and use (Warwick 2008a)
5. completion and sustainability (Brown et al. 2009; Kretzschmar 2009; Sewell 2009)
6. digital humanist career (Siemens et al. 2010).

These topics were addressed as a series of questions which participants answered reflecting on their particular experience and the projects they had worked on.

3. Results and Discussion

The first workshop had fifteen participants and the second twelve. Table 1 shows a breakdown of participants by subject. Examples of the types of DH projects carried out by the participants were digital collections and libraries (modern short novels and poetry, XIX century manuscripts marked in TEI), linguistics (text mining, corpus building, corpus of Mexican Spanish), digital images (research in pre Hispanic mural paintings, visualization archeological sites) and Anthropology (sound files for linguistics research indigenous languages). Almost all participants were project leaders, with the exception of two programmers, one MA publishing student and one graphic designer.

Subject	Participants workshop 1	Participants workshop 2
Anthropology	1	0
Architecture	0	1
Art History and Aesthetics	3	3
Bibliography and Book Studies	2	3
Engineering Linguistics	1	0
History	1	1
Philology and Literature	3	3
Philosophy	4	1

Table 1

Of the people attending only a few were aware of the field of DH. Some had links with international projects (such as Biblioteca Virtual Cervantes) but in general most work was local and individual. Participants were pleased to discover "that there are other people like me".

Almost all projects are personal and not institutional initiatives. Several scholars remarked that university authorities had a vague idea of the importance of registering and managing digital materials but that this institutional support in practice lacked coherent policies or structures for them to have any real impact. None of the projects for example, had any specific physical location assigned and many had to improvise working spaces in order to cope with human resources and equipment.

A few people remarked however that working marginally was actually a good thing. "Being ignored can also be an advantage. Being independent and invisible to the institution allows you to be dynamic and creative".

Surprisingly, funding was not a problem for any of the projects. Most scholars applied and received funding from government or the UNAM. However, quantities were not particularly large (around 15,000 US dollars) and no project had permanent institutional funding. Participants did not mention a lack of access to computational technologies which had been suggested as a possible problem for DH projects in developing countries (Terras 2006).

Projects were rarely documented with the exception of Linguistics. This is similar to other DH projects worldwide (Warwick 2009). Due to one to three year funding periods many felt that they were in a race against time to complete and documentation was left out.

Finding, training and retaining human resources are also key issues. All participants agreed that it is difficult to find human resources with the necessary skills and training was required. Additionally participants themselves went through a steep learning process whilst developing their project and found little learning support. In terms of training in Mexico there are no DH centres or courses. However, a couple of MA DH related classes are in development. Sharing information and pooling resources was considered fundamental towards developing the field. Participants noted the urgent need to compile best practice and guidelines as there seemed to be duplication of efforts and no communication.

Long term sustainability of resources is a major issue. For example, many scholars had purchased their own server to host the project as there are no university guidelines for hosting projects. However, it is not clear what will happen once the servers have to be replaced, or if the researcher left the university. In other cases, projects were hosted at the Computing Services department but usually as a personal rather than a formal collaboration. Others have hosted their projects on external servers, sometimes even at their own personal expense. As one participant remarked "when does a project become a university service and therefore somebody else's responsibility?" This is a common issue for DH (Brown et al. 2009; Kretzschmar 2009; Sewell 2009). We detected a notable absence of the library community whose skills are essential to these issues. Participants were aware of preservation but had not addressed the issue at all.

Another major issue was evaluation and recognition of DH work. Many felt that their work, although funded, was later not taken into account for evaluation purposes. However, it was also noted that it is difficult for evaluation committees who have no experience or knowledge about these types of projects to assess them. Many had worked individually with their departmental boards but it was agreed that providing tools, acting as a consultative body and lobbying collectively would be a more effective approach.

4. Conclusions

Results from the workshop indicate that forming a DH community is possible as we found sufficient projects, scholars and interest to sustain a working group. All participants were enthusiastic about forming part of a local DH group. Workshops were by invitation only but have since resulted in other DH scholars coming forward and wanting to participate. Initial results indicate that issues and challenges regarding DH projects are similar to other countries and collaboration would be possible and fruitful. However, with some issues such as university and governmental recognition and support, guidelines and best practices and community awareness there appears to be a significant lag behind other countries. One main difference is the almost complete absence of the library community and this issue should be addressed. Main challenges are now: to discover and register more research and projects; develop best practices and guidelines in Spanish; incorporate the library community, build a directory of DH scholars; expand the group and develop mechanisms to increase national and international collaboration. In the next few months we will continue to work on more specific actions and report on them in due course.

References

Borgman, C. (2009). 'The Digital Future is Now: A Call to Action for the Humanities'. *DHQ: Digital Humanities Quarterly.* 3(4). http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

Brown, S. et al. (2009). 'Published Yet Never Done: The Tensions Between Projection and Completion in Digital Humanities Research'. *DHQ: Digital Humanities Quarterly.* 3(2). http://digitalhumanities.org/dhq/vol/3/2/000040/000040.html.

Clement, T., Jannidis, F. & McCarty, Willard (2010). 'Digital Literacy for the Dumbest Generation - Digital Humanities Programs'. *DH 2010 Conference Abstracts.* King's College London, 2010, pp. 31-39. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-815.html.

Dunning, A. (DATE). 'The Tasks of the AHDS: Ten Years On'. *Adriadne.* 48. http://www.ariadne.ac.uk/issue48/dunning/.

Friedlander, A. (2008). 'Asking Research Questions and Building a Research Agenda for Digital Scholarship'. *Working Together or Apart: Promoting the Next Generation of Digital Scholarship, Report of a Workshop Cosponsored by the Council on Library and Information Resources and The National Endowment for the Humanities.* Washington, D.C.: Council on Library and Information Resources, pp. 1-15. http://www.clir.org/pubs/reports/pub145/pub145.pdf.

Juola, P. (2008). 'Killer Applications in Digital Humanities'. *Literary and Linguistic Computing.* 23(1): 73 -83.

Kretzschmar, W.A. (2009). 'Large Scale Humanities Computing Projects: Snakes Eating Tails, or Every End if a New Beginning?'. *DHQ: Digital Humanities Quarterly.* 3(2). http://digitalhumanities.org/dhq/vol/3/2/000038/000038.html.

Muller, C. et al. (2010). 'The Origins and Current State of Digitization of Humanities in Japan'. *DH 2010 Conference Abstracts.* King's College London, 2010. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-630.html.

Presner, T. & Johanson, C. (2009). *The Promise of Digital Humanities. A White#paper..* UCLA.

Rehm, G. et al. (2007). 'Digital Text Resources for the Humanities - Legal Issues'. *Digital Humanities 2007 Conference Abstracts.* Urbana Champaign, 2007. http://www.google.com/url?sa=t&source=web&cd=1&ved=0CBYQFjAA&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.129.2070%26rep%3Drep1%26type%3Dpdf&rct=j&q=Digital%20Text%20Resources%20for%20the%20Humanities%20-%20Legal%20Issues&ei=r2t-TdeeEIGz0QGftZjTAw&usg=AFQjCNGLlTPhbaWWtN1f49Z6PmkzqqMZ2g&cad=rja.

Sewell, D. (2009). 'It's For Sale, So It Must Be Finished: Digital Projects in the Scholarly Publishing World'. *DHQ: Digital Humanities Quarterly.* 3(2). http://digitalhumanities.org/dhq/vol/3/2/000039/000039.html.

Svensson, P. (2010). 'The Landscape of Digital Humanities'. *DHQ: Digital Humanities Quarterly.* 4(1). http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

Terras, M. (2006). 'Disciplined: Using Educational Studies to Analyse 'Humanities Computing''. *Literary and Linguistic Computing.* 21: 229 - 246.

van den Huevel, C. et al. (2010). 'Building the Humanities Lab: Scholarly Practice in Virtual Research Environments'. *DH 2010 Conference Abstracts.* King's College London, 2010. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-611.html.

Warwick, C. et al. (2009). 'Documentation and the users of digital resources in the humanities'. *Journal of Documentation.* 65(1): 33-57. http://URL.

Warwick, C. et al. (2008a). 'If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data.'. *Literary and Linguistic Computing.* 23(1): 85 -102.

Warwick, C. et al. (2008b). 'The master builders: LAIRAH research on good practice in the construction of digital humanities projects'. *Literary and Linguistic Computing.* 23(3): 383 -396.

CloudPad – A Cloud-based Documentation and Archiving Tool for Mixed Reality Artworks

Giannachi, Gabriella
g.giannachi@exeter.ac.uk
Centre for Intermedia, University of Exeter

Lowood, Henry
lowood@stanford.edu
Stanford Libraries, Stanford University

Rowland, Duncan
dar@cs.nott.ac.uk
Mixed Reality Lab, University of Nottingham

Benford, Steve
sdb@cs.nott.ac.uk
Mixed Reality Lab, University of Nottingham

Price, Dominic
djp@cs.nott.ac.uk
Mixed Reality Lab, University of Nottingham

This paper reflects on the process of designing and building a documentation and archiving tool named CloudPad on the basis of its first evaluation at Stanford Libraries and the San Francisco Art Institute in September 2010. The paper explores the value of CloudPad and its ability to document individual users' replay of an artwork within the context of performance documentation and new media archiving, speculating on its possible use within a number of curatorial, educational and creative contexts that are relevant to digital humanities.

The CloudPad was developed in 2010 by a team in Horizon RCUK-funded digital economy research and involved staff in performance studies and computer science from the Universities of Exeter and Nottingham, with partners from Stanford Libraries, the Ludwig Boltzman Institute Media.Art.Research, The San Francisco Art Institute, British Library, Blast Theory, and the University of Sheffield. The work developed out of the team's intention to research novel theoretical and practical approaches for the documentation and archiving of mixed reality performances and artworks that span both digital and physical entities (Benford and Giannachi 2011), allowing users to engage with the materials creatively over time and from different locations. The project benefitted from previous research conducted by

members of the team through the AHRC-funded Presence project (2004-9), which used second life and a wiki to document practices spanning from performance art, to video art and new media, including work in virtual reality CAVE, and the EPSRC-funded Creator project (2008-9) which used an e-science tool, the digital replay system, to generate synchronised annotations about a mixed reality performance (DRS). The project also benefitted from the findings of the e-dance project (2007- 9), which was jointly funded by AHRC, JISC and EPSRC, and conducted by colleagues from the Universities of Bedfordshire, Leeds, Manchester and Open University. This adopted access grid technologies for developing new approaches to choreographic composition, involving the use of the Memetic toolkit for recording, replaying and annotating sessions in access grid. Finally, the project was developed in dialogue with artworks such as Lynn Hershman Leeson's RAW/WAR feminist film archives (2010), sosolimited's interactive archival performances, and current thinking in new media documentation (e.g. Costello 2005, Depocas et al 2003, Jones and Muller 2008 and Dekker 2011, among others).

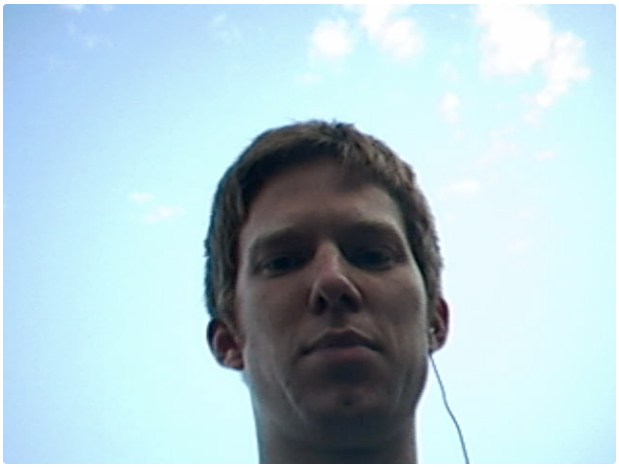
Technically the CloudPad was designed as a customisable web-based platform aiming to facilitate the synchronised playback and mash-up of cloud-based media entities such as video or audio files, as well as webpages and photographic materials, together with layers of user annotations. It took a novel approach to the archiving and replay of pervasive media experiences by making use of Web 2.0 technologies (DiNucci 1999) rather than grid technologies. CloudPad users were empowered to view the repository as a living document in which they could leave their own impression of an experience (both of the original event recordings as well as any thematic connections or annotations provided by other visitors and subject experts). Previous interactive systems designed for the replay of events for analysis lack this level of emergent reflection (see Brundell 2008), treating the corpus of recorded material as essentially immutable. To enable this, the CloudPad made use of internet-based storage, which means that media from a wide variety of different sources could be included in a presentation (for example YouTube videos can be included and synchronised with images from Flickr). This was accomplished by the use of HTML5 (see w3.org), an emerging web standard that enables collaborative interactive applications to be developed which run inside a web browser (Murray 2005).

As an initial form of content to assess the operation of the CloudPad we utilised a 'bespoke' documentation of Blast Theory's Rider Spoke that was recorded by our team when the work occurred at the ars electronica festival in Linz in 2009. Rider Spoke is a location-based game for cyclists developed by Blast Theory in collaboration with Mixed Reality Laboratory at the University of Nottingham as part of the European research project IPerG. The work encouraged participants to cycle around a city in order to record personal memories and make statements about their past, present and future that were associated with particular locations (see figure 1).

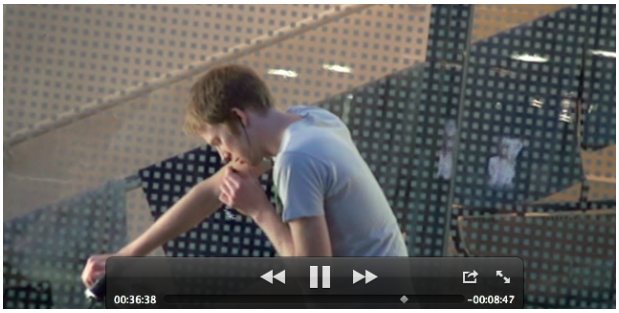
Blast Theory, Rider Spoke. Participant listening to recordings. Copyright Blast Theory.

To collect a documentation that addressed the complexity of this work, we developed a hybrid approach. This included the collection of documentations pertaining to the artists and technologists' descriptions of the works (in terms of original aims, interim analyses and final evaluations), as well as documentations of the user experience (see Jones and Muller 2008 and Depocas et al 2003), the latter recorded from a variety of points of

view (e.g. first person, third person) and through a number of technologies (e.g., video, GPS, Wi-Fi) and perspectives (see figures 2, 3 and 4).



Linz documentation. Participant captured via first person point of view.



Linz documentation. Participant captured via third person point of view.

Linz documentation. Participants journey through the city captured on googlemaps.

The overall analytical approach was interdisciplinary, thus including different and potentially even contrasting accounts of the event (see Chamberlain et al 2010). These accounts were presented through a number of historic, canonic and participant 'trajectories' (see figure 5).

Matt Adams' annotation about a participant linking first and third person perspectives in a canonic trajectory.

By historic trajectories we defined a historic event, i.e. a participant's experience as documented in a video; by canonic trajectories we defined an expert user's set of annotations through these materials; and by participant trajectories we defined the CloudPad user's own annotations (Giannachi et al 2010). This architecture does not privilege a single viewpoint and encourages creative use of both the historic materials and their canonic annotations. Arguably, every replay, producing participant trajectories, re-constitutes the work.

The CloudPad evaluation showed that users did not only envisage adopting the CloudPad for purposes of documentation and archiving, but also wanted to use it curatorially, to present work to others and engage users in annotating materials, for example in an online exhibition, academically, to write 'visual essays', and creatively, to make artwork. We have seen that the CloudPad offers scholars, artists and students the possibility to document, archive, curate and create synchronised variable media mash-ups from existing digital resources. These mash-ups, which show how users have engaged with the original documentation stored on CloudPad, build an invaluable resource for those who may be interested in how a core documentation or archive is navigated and interpreted over time. In other words, the CloudPad is not only a documentation and archiving tool, it also documents and archives itself, generating contextual footprints or traces and possibly even re-enactments of every replay of the original materials. This paper reflects on the advances generated by this particular functionality in terms of performance documentation, preservation, and re-enactment.

2. Acknowledgements

We gracefully acknowledge the RCUK funded Horizon digital economy research and the AHRC funded Riders Have Spoken project. We would like to thank Blast Theory and staff at the Ludwig Boltzman Institute

Media.Art.Research, Katja Kwastek, Dieter Daniels and Ingrid Spoerl in particular, who facilitated the documentation of *Rider Spoke* in Linz, and our participants and volunteers who gave their time to make this documentation possible. We would also like to thank the staff and students at the San Francisco Art Institute, Stanford Libraries and St Jose State University for providing crucial feedback that informed the writing of this paper.

References

Benford, S. and Giannachi, G. (2011). *Performing Mixed Reality.* Cambridge, Mass.: the MIT Press.

Brundell, P., Tennent, P., Greenhalgh, C., Knight, D., Crabtree, A., O'Malley, C., Ainsworth, S., Clarke, D., Carter, R. and Adolphs, S. (2008). 'Digital Replay System (DRS) - a tool for interaction analysis'. *Proceedings of the 2008 International Conference on Learning Sciences.* Utrecht: ICSL, June 23-24, 2008.

Chamberlain, A., Rowland, D., Foster, J., Giannachi, G. (2010). 'Riders Have Spoken: Replaying and Archiving Pervasive Performances'. *Leonardo Transactions.* 43.1: 90-1.

Costello, B., Muller, L., Amitani, S., and Edmonds, E. (2005). 'Understanding the Experience of Interactive Art: Iamascope in beta_space'. *ACM 2005.* vol. 123: 49-56.

Dekker, Cosetta Saba, Julia Noordegraaf, Barbara Le Maître and Vinzenz Hediger (eds.) (forthcoming 2011). *Preserving and Exhibiting Media Art: Challenges and Perspectives.* Amsterdam: University Press.

Depocas, A., Ippolito, J., Jones, C. (2003). *Permanence through Change: The Variable Media Approach.* New York: Guggenheim Museum Publications.

DiNucci, D. (1999). *History Fragmented Future Recovered.* `http://www.cole20.com/web-20-history-fragmented-future-recovered/` (accessed October 2010).

Giannachi, G., Rowland, D., Benford, S., Price, D. (2010). 'The Documentation and Archiving of Mixed Media Experiences: the Case of Rider Spoke'. *Digital Futures.* Nottingham, 11-12 October 2010.

Horizon. `http://https://www.horizon.ac.uk/` (accessed October 2010).

IPerG. `http://iperg.sics.se/index.php` (accessed October 2010).

Jones, J. and Muller, L. (2008). 'Between Real and Ideal: Documenting Media Art'. *Leonardo.* 41.4: 418-41.

Murray, G. (2005). *Asynchronous JavaScript Technology and XML (Ajax) With the Java Platform.* `http://www.oracle.com/technetwork/articles/javaee/ajax-135201.html` (accessed October 2010).

Moving Beyond Anecdotal History

Gibbs, Fred
fwgibbs@gmail.com
George Mason University

Now almost fifty years old, Walter Houghton's seminal work, *The Victorian Frame of Mind, 1830-1870*, has influenced generations of scholars of the nineteenth century and remains the primary introduction to Victorian thought that every student in the field reads. From a close reading of famous Victorian writers such as John Stuart Mill and Thomas Carlyle, Houghton argued that the Victorians were characterized by specific, common personality traits such as optimism, hero worship, and earnestness. Houghton believed that these traits were visible in the rise (or decline) in the use of particular words and phrases, such as an increasing use of "light," "sunlight," and "hope" as illustrative of their optimistic world view.

Despite the enormous impact of *The Victorian Frame of Mind* on generations of scholars across the humanities, it has not been accepted uncritically. Many concerns stem from Houghton's myopic textual methodology: generalizing the character of a people —"the Victorians"—from the words of a select few. Although Houghton cites hundreds of primary sources in his bibliography, his book has been characterized as anecdotal, elite intellectual history. Despite such criticisms, Victorianists have been able neither to thoroughly assess the general validity of Houghton's theses nor to offer alternatives.

New digital tools and the vast digital library of Google Books now allow us to conduct a comprehensive survey of Victorian writing—not just the well-known Mills and Carlyles, but tens of thousands of lesser-known or even forgotten authors—to test whether the Victorians truly did use the kinds of words and phrases that Houghton claimed they did. Did metaphors of hope *actually* increase in real terms between 1830 and 1870? Or was this only true for the dozen prominent writers he examined for his chapter on optimism? How can we complicate Houghton's characterizations and understand change over time through the vast index of Google Books? Can we refine the timeline for the emergence of his characteristics, moving beyond the disturbingly neat, rounded-year boundaries he set for his book? How can we correlate historical events with disturbances in the linguistic data? To what extent can we separate cultural history from printing history with a large corpus of digitized literature?

Dan Cohen, my colleague at the Center for History and New Media, and I have begun work on a project to answer these very questions. With the help of a Google Digital Humanities Grant, we're attempting what Franco Moretti calls a "distant reading" of the Victorians. My paper will explain how we've gone about querying the data available through Google Books, how we've been able to make sense of and interpret the results, and how we've dealt with the messiness of the data. I hope to solicit feedback about our methodologies and conclusions as part of a larger discussion about how the Google Books corpus (and similar datasets) could be made more usable for large-scale data mining projects relevant to the diverse research interests of the audience. How far can we push our methodologies beyond testing certain theses in order to allow the texts to speak for themselves?

In terms of the Victorians, some preliminary results have proven quite intriguing. For example, the number of books published with "universal" in the title declined steadily throughout the century, but earlier than most interpretations in the secondary literature point out. A look at published titles suggests that the terms "God," "Christian," and "Bible" follow rather different contours, though explanations are not immediately apparent. Similarly, how can we explain the striking publication parallels between the terms "belief" and "Aristotle"? The median number of titles that use the word "hope" does not significantly change between 1830 and 1870: does this cast some doubt on Houghton's characterizations? Or does it simply indicate that book titles are not an accurate gauge of popular sentiment? We have not yet been able to examine the full texts from the publications that are being counted, but we hope to do that soon. To what extent will a more sophisticated linguistic analysis of the full texts reinforce or contradict what the titles alone tell us? Does this have implications for similar large-scale research methodologies?

Historic Interpretation, Preservation, and Augmented Reality in Falmouth Jamaica

Graham, Wayne
wayne.graham@virginia.edu
University of Virginia

Nowviskie, Bethany
bethany@virginia.edu
University of Virginia

Despite the ground-breaking work of graphics visionaries like Alan Sutherland[1] and Ed Catmull[2] in the 1960s and 1970s, which unlocked computer screens as interactive tools, hardware portability issues have constrained computer interaction to a two-dimensional space which often simulates the real world. However, the considerable market growth of sophisticated mobile devices over the last several years has begun to push the boundaries of interaction in virtual environments. Instead of experiencing a simulated environment sitting at a computer, users are shifting to experiencing physical environments with a computing device capable of enriching their subjectie experience of the space. The success of augmented reality systems like BionicEye, RobotVision, TATAugmented ID, and Layar provides glimpses at how this technology might be leveraged by cultural heritage institutions, individual academics, and even local, municpal officials, to provide opportunities for students and the general public to interact with space and place in new and exciting ways.

The University of Virginia's Department of Architectural History holds its training-oriented field school, under the direction of Professor Louis Nelson, in the city of Falmouth Jamaica each year. Falmouth is a fruitful city for study because of its unique history, which makes it the best preserved example of Georgian architecture in the Caribbean. Founded in 1769, the city was originally designed as the main northern port for the island's thriving sugar trade. Tied closely to the infamous "Triangle-trade", the slave-based sugar economy allowed the city to grow until the emancipation of slaves in the British Empire in 1840. After 1840, the town saw a significant decline and experienced very little construction over the next 170 years, preserving its architecture and urban design as an early nineteenth-century time capsule.

In 2009, the Jamaican government approved development of a cruise ship terminal in Falmouth. Its first ships are scheduled for docking in the summer of 2011. The predicted influx of tourism to the town will be an economic boon to residents, but will also bring significant changes to the architectural identity of the town. The 11-acre port is of such size that, while ships are docked, they will tower over the Trewlany Parish Church of St. Peter, the tallest building in the town. New dining and recreation facilities will be constructed around the terminal, and a simplified interpretation of early nineteenth-century life will be constructed for the enjoyment of tourists. This will result a better standard of living for most residents of the town, but will mean the loss of an historical laboratory for architecture. As businesses sprout up, residents will be pushed away from areas around the terminal and century-old houses will be torn down and replaced with store-fronts.

Realizing that the introduction of the cruise terminal to the city will forever alter the architectural identity of the town, the UVa Architectural History Field School recently shifted its focus from deep analysis of a handful of buildings in a single summer, to a more general effort to survey the status of the buildings of the entire city. This survey includes measured sketches, colorcoding of a given structure's overall condition, images of the structure as it stands today, and information on when the building was constructed, construction material, and other items of interest to architectural historians. These findings were then submitted to UVa Library's Special Collections for long-term archiving.

Sensing a real opportunity to provide access to this important work, the Scholars' Lab partnered with Louis Nelson, chair of the Architectural History department, to investigate ways to make this nearly decade's worth of research available to a wider audience. We identified three groups of users for the content: academics interested in the underlying data of the city, government officials who need to plan city restoration efforts, and tourists interested in finding out more about the town they will visit. The Scholars' Lab was particularly interested in interface design decisions that would serve each of these communities well.

In order to provide different mechanisms of access for three distinct user groups, we employed several open-source tools to create a solution that would expose this valuble architectural and historical data to a large and varied audience. Leveraging our own expertise in open-source Geographic Information Systems infrastructure[3] with the flexibility of the Omeka collections and exhibits framework, the Scholars' Lab

has constructed several different ways in which these different audiences may interact with library-curated spatial data.

In order to allow city planners access to the underlying data, maps originally drawn in AutoCad were converted to a GIS format and loaded on to a web-accessible server. Web services were created to allow high-end users of GIS software access to the information with proprietary tools like ArcGIS. We then utilized those same web services to create map interactions within Omeka using the open source mapping library OpenLayers. Images and metadata (along with the full architectural survey report) were uploaded to Omeka and a custom VRA Core metadata standard was created to allow more appropriate description of the architectural elements in the collection. Long-form academic essays are also in the process of being written for approximately 30 of the most important structures in the city. These will be presented along with maps highlighting where important structures are located, to allow visual methods for browsing the collections to function in tandem with scholarly interventions to highlight specific structures of interest. Importantly, this approach also allows us to expose underlying geographic information as web services, allowing other scholars to reuse the information.

While the site boasts other advanced features, including a faceted Solr-based search, we also wanted to explore methods for exposing the same data in new ways to the different target groups for the project. With the growth of web-enabled smart devices, we plan to try two experiments at the Vernacular Architecture Forum's Annual Meeting, in June 2011 in Falmouth, Jamaica. The first will place QR codes at selected structures around the town to allow individuals with web-enabled phones to access all of the information about the structure presented on the Scholars' Lab website. We are also building a Layar-based augmented reality browser for the city which will allow individuals with smart phones (iPhones and Android-based devices) to install a simple application that will overlay information about buildings on a viewport, accessed simply by pointing their phones at the building.

This is an ongoing project, with field experiments and user feedback scheduled to be conducted before the Digital Humanities 2011 conference. We hope to model a technical approach to providing access to library-curated information for multiple audiences, using different technological approaches and techniques to frame the data not only in terms of scholarly use, but for tourism and cultural heritage appreciation, and for the more practical employment of the data by city planners to optimize restoration efforts. We also hope that these tools will help raise awareness about the fragility of the town to tourists, and can act as a way to expose UVa students' and scholars' research in multiple formats to those interested in underlying historical and spatial data, and the multimedia and embodied arguments which have been crafted using new tools and methods.

Notes

1. Alan Sutherland built the first graphics program named "Sketchpad" as part of his PhD. work in 1963. Sketchpad allowed users to draw geometric shapes and create copies of them. This work was not only important in the realm of computer aided design (it is viewed as the grandfather of modern CAD software), but also in its organization of the code in objects, which is the basis for modern object oriented design.
2. Beyond his current role as CTO of Pixar, Catmull is perhaps one of the most important people in the world of computer graphics, discovering methods to apply images to geometries, smoothing lines drawn on computer screens, among many others. He was also a student of Sutherland...
3. In 2009 and 2010, the Scholars' Lab hosted an NEH-funded "Institute for Enabling Geospatial Scholarship:" `http://lib.virginia.edu/scholarslab/geospatial/`. At DH 2009, several members of the Scholars' Lab presented a panel discussion ("New World Orderings") on GIS, including a description of our open source , web services-based Geospatial Data Portal: `http://gis.lib.virginia.edu`. And at DH 2010, Scholars' Lab director Bethany Nowviskie presented a poster on possibilities for spatial humanities, "Inventing the Map."

The Digital Materiality of Early Christian Visual Culture: Building on John 20:24-29

Heath, Sebastian
sebastian.heath@nyu.edu
Institute for the Study of the Ancient World, New York University

This paper explores the nature of digital materiality as it resides in the visual and written culture of Early Christianity. Within archaeology and related disciplines, “materiality” is a theoretical approach that focuses on physical things - such as objects, books, or buildings - as one starting point for building an understanding of past thought and behavior (White 2009). As a term, "digital materiality" does not yet have a fixed meaning and can refer to the physical manifestations of the computer age (Manoff 2006), to the processes by which digital representations become physical architecture (Gramazio and Kohler 2008) or to the effects of digital information in the modern world (Leonardi 2010). Here, I mean "digital materiality" as the transport of information about the material culture of past societies, and particularly the material culture of Early Christianity. Looking for fluid relationships between thought and object in ancient evidence suggests that "digital materiality" is an appropriate metaphor for both recovering past interactions with material culture and for describing the role of networked information in modern archaeological and art historical scholarship. It is this intersection of past and present that is of particular interest. While stressing potential, this paper also looks to the practical consequences of current efforts to digitize ancient activity that survives in material form.

Existing virtual representations have already exposed clear overlaps between the written word as object and the manifestation of those concepts in visual media. The Codex Sinaiticus is a fourth century codex bible removed from Saint Catharine’s monastery on the Sinai Peninsula in Egypt in the 19th century and now largely in the British Museum. Its early date makes it plausibly the first extant bible as that term is conceived in Christian terms. Most of the surviving pages of the codex are available online at the site http://codexsinaiticus.org/ . Among the passages found there is John 20:24-29, where the disciple Thomas, of “doubting Thomas” fame, demands to touch the wounds of the resurrected Jesus when he appears to his followers. The passage ends with the exhortation, “20:29. Because you have seen me, you have believed; blessed are those who have not seen and yet have believed.”

The traction this concept had in Early Christian culture is clear from a hammered gold disk produced in Egypt and now in the collection of the American Numismatic Society that has a stable digital representation available at the URI http://numismatics.org/collection/0000.999.51006 . This physical object quotes the text of John 20:29 while illustrating Thomas in the act of touching Christ. This paper stresses that the modern opportunity to engage in such a self-referential illustration of the materiality of thought in the Later Roman Mediterranean is a serendipitous result of independent efforts to digitize the material record of that time and place. Just as the creation of the surviving material record should be recognized as the cumulative action of many individuals, it is likely that exploration of that record will be enabled by many projects and institutions working within their own areas of expertise and with content specific to their domain (Heath 2010, Terras 2010). It is the interactions of a series of self-digitizing and independent communities – here Early Christian textual studies and Numismatics – that can recover relationships between physical object and human thought that is a primary goal of materiality as a methodological approach.

It is, of course, important to recognize that while the Internet will make evident the material implications of past human thought and action, it will not of its own bring scholars into direct contact with the material culture they study. Digital Humanities as applied to archaeology and visual culture will usually mean working with surrogates: one cannot download an object, one can only see its representation . The network does not take us to a site, it only provides access to descriptions and pictures. Digital materiality is therefore an act of transmission (Liu 2004) so that its deficiencies leave it open to criticism.

Trends within the study of textual evidence as embodied in manuscripts suggest that this observation is not a barrier to analytical progress. Projects such as the Codex Sinaiticus digitization effort are showing that digital access to manuscripts is returning the material to a central place in the study of primary sources that had been abstracted in critical editions. The digitized page images show in great detail the large number of variants and corrections that make plain that written evidence for the ancient world does not exist independently of its physical media. The Homer Multi-Text project (http://chs.harvard.edu/

chs/homer_multitext) is self-consciously engaged in enabling virtual access to multiple manuscripts of the Iliad and Odyssey that range from the Hellenistic to Medieval periods. Such initiatives indicate that material and thought become meaningfully unified in a digital domain. Accordingly, digital materiality is not a poor substitute for direct autopsy of material culture. Rather than de-emphasizing the physical, Digital Humanities will bring it to the fore. But it needs to be pointed out that at this moment, the best critical editions of the *New Testament* (Aland and Nestle 2006, Aland et al. 2006), with rich apparatus for the Gospel of John, is not available online so that commercial interests are an impediment to to the study of the text, whether considered as an idealized abstraction or a material object. This suggests that the constituent components of both modern and ancient digital materiality are at a transitional point where "primary sources" are more accessible than "secondary works".

It is particularly important to stress this point when we recognize that it is no longer possible within archaeological scholarship to have hands-on access to all relevant material (Stewart 2008). To recall the ending sentiment of the doubting Thomas story, “blessed are those who have not seen and yet have believed.” This can be applied to the current state of archaeological and material studies and lightly reformulated as an invitation to both make use of the full potential of material and textual sources on the internet, and to aggressively pursue such availability.

References

Aland, B. et al. (2006). *Greek New Testament 4th Edition.* Peabody, MA: Hendrickson Publishers.

Aland, B., Nestle, E. (2006). *Novum testamentum Graece 27th Edition.* Stuttgart: Deutsche Bibelgesellschaft.

Gramazio, F., KOHLER, M. (2008). *Digital Materiality in Architecture.* Baden: Lars Müller Publishers.

Heath, S. (2010). 'Diversity and reuse of digital resources for ancient Mediterranean material culture'. *Digital Research in the Study of Classical Antiquity.* G. Bodard, S. Mahony (eds.). Farnham, UK: Ashgate, pp. 35-52.

Leonardi, P. (2006). 'Digital materiality? How artifacts without matter, matter.'. *First Monday.* 15.6 (6-7 Jine 2010). http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3036/2567.

Liu, A. (2004). 'Transcendental data: toward a cultural history and aesthetics of the new encoded discourse'. *Critical Inquiry.* 31 (Autumn 2004): 49-84. http://criticalinquiry.uchicago.edu/features/artsstatements/arts.liu.htm.

Manoff, M. (2006). 'The materiality of digital collections: theoretical and historical perspectives'. *Portal: Libraries and the Academy.* 6.3 (July 2006): 311-235. http://dspace.mit.edu/handle/1721.1/35689.

Stewart, P. (2008). *The social history of Roman art.* New York: Cambridge University Press.

Terras, M. (2010). 'Digital curiosities: resource creation via amateur digitization'. *Literary and Linguistic Computing.* 25.4 (2010): 425-438. http://llc.oxfordjournals.org/content/25/4/425.full.

White, C. (2009). *The materiality of individuality: archaeological studies of individual lives.* New York: Springer Science.

Image Markup Tool 2.0

Holmes, Martin
mholmes@uvic.ca
Humanities Computing Media Centre, University of Victoria

Timney, Meagan
mbtimney@uvic.ca
Electronic Textual Cultures Laboratory, University of Victoria

This paper discusses the re-development of The Image Markup Tool in two parts: (1) as a crossplatform desktop application and (2) as a web-based, html5 standard, client-side browser application. Currently, a few *text-based* tools allow for markup of documents (most often) in XML/TEI. They range from legacy software such as the Analytical System Tools and SGML/XML Integration Applications (Anastasia)[1] and Editing Digital Interactive Texts in an Online Network (EDITION),[2] to more recent projects such as eLaborate,[3] TextGrid,[4] and TEXTvre.[5] TextGrid uses a collaborative document markup interface, as well as a project and user management system to facilitate the markup of texts. TEXTvre, which will be modeled on TextGrid, has been described as a "a working exemplar VRE for textual scholarship." While each of these tools offers specific methods of text-based editing, The Image Markup Tool is *image-based*, allowing for the markup of encoded digital images of remediated textual objects.

We are building on the current iteration of The Image Markup Tool (v1), developed by Martin Holmes at the University of Victoria. Version 1 of the Image Markup Tool was first written in 2006, and has gone through several versions. It was originally conceived as part of the project *Le mariage sous L'Ancien Régime*,[6] where it was used to mark up engravings (Carlin, Haswell and Holmes 2006; Carlin and Holmes 2008). Early versions of the tool used SVG code embedded into a TEI file to delineate areas of interest on an image, but the current version makes use of the Facsimile module recently added to the TEI schema, which enables the use of native TEI elements to define rectangular "zones" on "surfaces" (often pages) which are part of a facsimile.

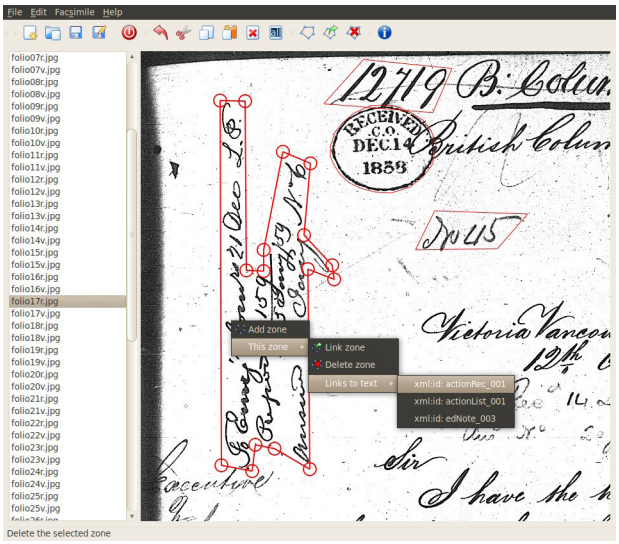
However, version 1 of the Image Markup Tool suffers from a number of serious limitations, most of which were apparent from its inception. Firstly, it allows only rectangular areas (hereafter "zones") to be specified on images. This was in line with the original specification of the Facsimile module in TEI, in which <zone> elements were similarly constrained, but users have been demanding the ability to specify polygonal shapes ever since the first release, and recent modifications to the TEI schema now allow the use of polygonal <zone> elements. Secondly, IMT version 1 can handle only one image per file. This was sufficient for its original projected use as part of the *Mariage* project, which focused on individual engravings, but makes the tool inadequate for serious facsimile work; most documents have multiple surfaces or pages. Thirdly, the program can handle only a one-to-one relationship between a single <zone> on a <surface>, and a single in the <body> of the document. This is inadequate. It is a common requirement to link, for instance, a single block of text on an image to an original transcription, a modernized version of the transcription, and an editorial note or interpretation. Similarly, a single (or in fact any other element) in the body of a text might conceivably be linked to more than one <zone>; multiple views of a particular surface or page might be provided in the facsimile, each with an equivalent <zone>. Finally, IMT version 1 was written as a Windows application using Borland Delphi. It will run on Linux using Wine, but there is no simple way to run it on a Macintosh computer.

Nevertheless, IMT version 1 has a number of strengths. As a compiled desktop application, using a very sophisticated open-source graphics library (Graphics32), it can to do high-quality resampling of images on-the-fly, enabling effective and rapid zooming of high-resolution images. Most file operations are very fast, and the interface itself is simple and relatively easy to use. For Windows and Linux users, it is easy to download and install.

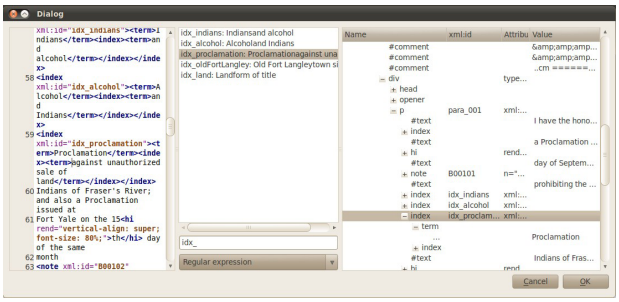
The desktop version of IMT 2 is being written using Nokia's QT Creator tools. This will enable us to compile the application for Windows, Mac and Linux. Our intention is to build on the strengths of version 1 -- in particular, the speed and efficiency of graphics handling and file i/o, and the user-friendliness and simplicity of the interface -- while adding three important improvements:

1. Handling of an unlimited number of images.
2. Many-to-many linking between <zone>s and any elements with @xml:id attributes in the <text> section of the file.
3. Support for polygonal zones.

These screenshots of early development pilots show how we envisage the user interface.



The first shows the main window, with multiple folios (pages) listed down the left side. Clicking on a folio shows the corresponding image, and the polygonal zones defined on that image (in red). The selected zone has draggable circular nodes at all of its corners. The context menu for the selected zone is displayed, showing that the zone is linked to three different elements in the <text>.



The second screenshot shows the corresponding dialog box, where the user can select an element in the <text> of the TEI file for linking to a zone. Three views of the file are available: the first is a read-only syntax-highlighted text view, the second consists of a list of all the @xml:id attributes in the file, along with a text box which can be used to filter them, and the third is an "outline" or tree view of the file. The user could use any of these to find and select an element to be linked to a zone. In addition, the user will be able to edit any element in the <text> of the TEI file directly, to add transcription, markup, @xml:id attributes, etc. We do not envisage that the majority of XML editing will be done in the Image Markup Tool itself; rather, the base transcription would be done using an XML editor such as oXygen, and the file brought into the IMT for the definition and linking of images and zones. However, it will be important to allow direct editing of the XML code so that corrections and changes can be made without moving the file back into an XML editor.

The web platform of the Image Markup Tool fills a gap in current collaborative editing models, and will provide a lightweight "edit-anywhere" version of the desktop application. The web-platform will be built for a large population of needs, but our first user-case study will be the Editing Modernism in Canada Project (EMiC).[7] Rather than following past practices of transcribing texts and marking up transcriptions in the creation of electronic texts, EMiC and its partners will pioneer image-based editing, semantic markup, analysis, and visualization of texts in a field of emergent practices in digital-humanities scholarship. Instead of producing reading environments based on linear-discursive transcriptions of texts, EMiC will produce in collaboration with its partners techniques and technologies for encoding and interpreting the complex relations among large collections of visual and audial objects in non-linear reading environments.

Our rationale for a web-based browser application includes facilitating RESTful architecture and interoperability with other systems via API (including, for example, Scripto,[8] developed at the Centre for History for New Media at George Mason University). The IMT web-platform will allow a user to load images and XML documents into a browser window (either locally or via URL). The drawing and linking of the polygonal <zone>s now supported in the TEI schema are made possible with the HTML5 <canvas> element. As in the desktop application, it is assumed that the user will perform most TEI markup with an XML editor (such as oXygen), but the application will also support lightweight XML editing. The most powerful feature of the web-based application will be the potential to feed linked XML documents and images directly into a collection-builder such as Omeka[9] to facilitate scholarly edition building and electronic publishing via a suite of tools.

References

Carlin, Claire, Eric Haswell and Martin Holmes (2006). 'Problems with Marriage: Annotating Seventeenth-century French Engravings with TEI and SVG'. *Digital Humanities 2006 Conference*. July 2006http://https://webcgi.oulu.fi/dh2006/viewabstract.php?id=17http://www.allc-ach2006.colloques.paris-sorbonne.fr/DHs.pdf.

Carlin, Claire and Martin Holmes (2008). 'Domestic strife in early modern Europe: images and texts

in a virtual anthology'. *Digital Humanities 2008.* 26 June 2008http://www.ekl.oulu.fi/dh2008/Digital%20Humanities%202008%20Book%20of%20Abstracts.pdf.

Timney, Meagan and Dean Irvine (Forthcoming: 2011). 'A New Build: Digital Tools for Archives, Commons, and Collaboration.'. *Archival Narratives for Canada: [Re] Telling Stories in a Changing Landscape.* Kathleen Garay and Christl Verduyn (ed.). Fernwood Publishing.

Notes

1. (ITSEE, University of Birmingham; http://www.sd-editions.com/anastasia/index.html)
2. http://www.sd-editions.com/EDITION/
3. Huygens Instituut KNAW (a research institute for text edition and textual scholarship of the Royal Netherlands Academy of Arts and Sciences); http://www.e-laborate.nl/en/
4. http://www.textgrid.de/
5. http://textvre.cerch.kcl.ac.uk/
6. http://mariage.uvic.ca/
7. http://editingmodernism.ca
8. http://www.scripto.org
9. http://www.omeka.org

The Tutor's Story: A Case Study of Mixed Authorship

Hoover, David L.
david.hoover@nyu.edu
New York University

The Victorian novelist and Christian Socialist Charles Kingsley (1819-1875) is now known mainly for his children's book, *Waterbabies* (Kingsley 1863), though he also wrote political and historical novels. Long after his death, his daughter, Mary St. Leger Kingsley Harrison (1852- 1931), discovered an unfinished and unexpected novel manuscript entitled "The Tutor's Story" among his papers. Mrs. St. Leger Harrison, writing under the name Lucas Malet, was herself one of the most famous novelists at the turn of the twentieth century, one who explored daring themes like incestuous desire, lesbianism, sadism, and prostitution (Schaffer 1996:109). Malet finished her father's novel and published it in 1916 (Kingsley and Malet 1916).

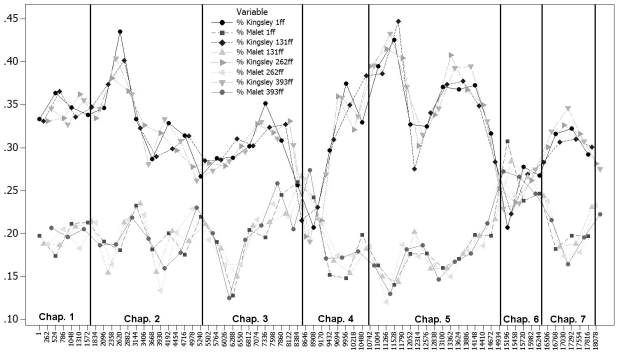
In her preface, Malet describes the state, size, and nature of the manuscript, and this description gives a fairly solid basis for assigning at least parts of the novel to the two authors. She tells us that the beginning of the manuscript, and so presumably also the novel, was "fairly consecutive," so that we can expect the early chapters to be Kingsley's. But she also tells us that there were other "chapters and skeletons of chapters" from much later in the story, without further indicating where these occur in the novel. Finally, she reports that the plot was unresolved, and that she doubled the size of the text in completing it (Kingsley and Malet 1916: vi). This suggests that the late chapters of the novel are probably mostly by Malet. This complex and difficult scenario provides a good opportunity for testing the effectiveness and limitations of some old and some new methods of authorship attribution, including t-tests, Burrows's Delta (Burrows 2002, 2003; Hoover 2004a, 2004b), and Craig's version of Burrows's Zeta (Craig and Kinney 2009; Hoover 2010).

As is so often true in the real world, some aspects of this authorship problem are not exactly what one would want. Some of Kingsley's novels are as much social commentary as fiction, dealing with issues like the plight of the rural poor, poor sanitation, child labor, and the exploitation of workers. Others are historical novels, set in Anglo-Saxon times, during the reign of Elizabeth I, and fifth-century Alexandria. Two others are children's books. Given this varied output, it is

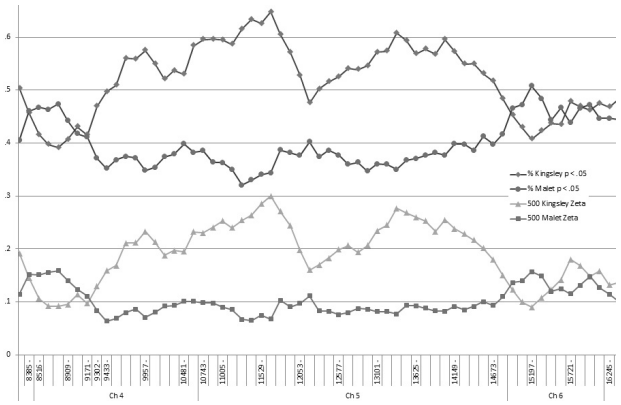
difficult to assemble sufficient similar Kingsley texts for testing. Furthermore, Malet tells us that she has tried to match her style to that of her father, and contemporary reviews of the novel comment that the book sounds just like Kingsley (Book Review Digest 1917). Finally, while most of Kingsley's fiction is third-person, this is a first-person novel. Malet's fiction is less varied, but it is also mostly third-person.

In spite of these difficulties, initial PCA, Cluster Analysis, and Delta tests on a group of novels by Kingsley and Malet all very successfully distinguish the two authors. Delta results remain quite accurate even for short sections, typically about 90% accurate for Kingsley, and often 100% accurate for Malet, even on large numbers of 500-word sections. Because we can expect some relatively short passages by each writer interspersed with passages by the other, it seems reasonable to test the entire novel divided into sections of 524 words (the novel divides almost exactly into sections of this size). In order to identify changes of authorship in such brief passages, the novel is tested with rolling segments of 524 words. The first section comprises the first 524 words; the next section comprises the 524 words that begin at word number 132, the next the 524 words that begin at word number 263, the next the 524 words that begin at word number 394, and so on through the rest of the novel. Rolling segments have been put to good use in several authorship attribution and stylistics studies; see Craig (1999), Burrows (2010), and van Dalen-Oskam and van Zundert (2007).

I am testing the rolling sections of the novel in three ways. The first uses a list of 2873 marker words that t-tests identify as being used significantly differently by the two authors ($p < .05$). The percentage of the word types (really individual spellings) in each section that belong to each author's set of marker words is graphed in Fig. 1. The upper set of lines show the percentages of Kingsley marker words and the lower set the percentages of Malet marker words. For example, in the first sections of Chapter 1, about 33% of the types are Kingsley marker words and about 18% are Malet marker words. The separation of the two sets is nicely distinct for the first three chapters, all of which, as expected, are attributed to Kingsley. The beginning of Chapter 4 seems to contain some of Malet's writing, and about the first two-thirds of Chapter 6 is attributed to Malet.



The t-test results for chapters 4-6 are repeated in a slightly different form in Fig. 2 (upper two lines; 20% has been added to the percentages for the t-test marker words to create a separation between the two sets of lines), along with results from Craig Zeta tests on the same sections (lower two lines). Rather than showing a separate line for each starting point, as in Fig.1, all the testing points for each set of marker words in Fig. 2 are joined by a single line. The graph for Craig Zeta shows the percentage of types in each section that are among the 500 most distinctively used Kingsley and Malet marker words. The smaller percentages for Zeta than for the t-tests reflect the fact that only 1000 marker words are used here, compared to the 2873 ttested marker words. Nevertheless, it is easy to see that Craig Zeta and t-tests give similar results and agree generally on the attribution of various parts of the chapters. Delta tests on similar-sized sections usually agree with these results as well. Many



Many of the chapters of the novel seem to be largely by one or the other author, but others seem thoroughly mixed. These results fall in line with what Malet's preface leads us to expect, and overall they seem fairly persuasive. A recent discovery makes them both more compelling and somewhat frustrating. After I had completed the testing described above, the problem seemed fascinating enough to deserve further research, and I began by trying to find out whether

Kinglsey's manuscript might still exist. Although I was not able to find any record of the manuscript, I came across a record of a copy of the novel in the Princeton Rare Books collection with Malet's penciled notes about which parts of the novel were written by Kingsley and which she wrote herself. For some chapters, her notes are quite precise, and they indicate that the attributions in Fig. 1 and Fig. 2 are essentially correct. For other chapters, she notes only that they are "mostly my father." Most frustrating of all is the fact that all markings cease after chapter 28 (of 41). The fact that the tests described above disagree with her notes for only 5-7 chapters suggest that, even texts involving mixed, joint, or collaborative authorship can be usefully investigated using these methods.

References

(1917). 'Kingsley, Charles. Tutor's Story.'. *Book Review Digest.* White Plains, NYOnline: Google Books: H.W. Wilson Company Volume 12.

Burrows, J. (2002). ''Delta': a measure of stylistic difference and a guide to likely authorship.'. *LLC.* 17: 267-287.

Burrows, J. (2003). 'Questions of authorship: attribution and beyond.'. *CHUM.* 37: 5-32.

Burrows, J. (2010). 'Never say always again: reflections on the numbers game'. *Text and Genre in Reconstruction. Effects of Digitalization on Ideas, Behaviours, Products and Institutions.* W. McCarty (ed.). Cambridge: Open Books.

Craig, H., and Kinney, A. (eds.) (2009). *Shakespeare, Computers, and the Mystery of Authorship.* Cambridge: Cambridge University Press.

Craig, H. (1999). 'Jonsonian chronology and the styles of A Tale of a Tub.'. *Re-Presenting Ben Jonson: Text, History, Performance.* M. Butler (ed.). Houndmills: Macmillan, pp. 210–32.

Hoover, D. (2010). 'Authorial style'. *Language and Style: Essays in Honour of Mick Short.* D. McIntyre and B.Busse (ed.). Palgrave.

Hoover, D. (2004). 'Delta prime?'. *LLC.* 19(4): 477-495.

Hoover, D. (2004). 'Testing Burrows's Delta.'. *LLC.* 19(4): 453-475.

Kingsley, C. (1863). *The Waterbabies: a Fairytale for a Land-baby.* London: Macmillan.

Kingsley, C. and Malet, L. (1916). *The Tutor's Story.* London: Smith Elder.

Schaffer, T. (1996). 'Some chapter of some other story: Henry James, Lucas Malet, and the real past of The Sense of the Past.'. *The Henry James Review.* 17.2: 109-128.

van Dalen-Oskam, K, J. van Zundert (2007). 'Delta for Middle Dutch—author and copyist distinction in Walewein.'. *LLC.* Vol. 22, No. 3.

Modes of Composition in Three Authors

Hoover, David L.
david.hoover@nyu.edu
New York University

I have already argued against the widely held belief that Henry James's switch from handwriting to dictation caused a radical change in his style (Hoover 2009). However, the wider question of how mode of composition affects literary style remains open, and James might be the exception rather than the rule. I report here on some preliminary studies for a more comprehensive examination of writers who changed their mode of composition either temporarily or permanently. The three authors examined here, Thomas Hardy, Joseph Conrad, and Walter Scott, present clear cases of changes from handwriting to dictation (and back), ones in which the details of composition are well known, and in which the changes in mode of composition take place within a single text.

The case of Thomas Hardy is slightly problematic because he was not always truthful about his wife's role in the production of his books, and in some cases burned some MS pages in her hand. For *A Laodicean*, one of his less important novels, however, the facts seem fairly clear. After sending the first three (of thirteen) installments of the serial version to the printer, Hardy fell ill, suffering from some kind of bladder inflamation. He struggled through the fourth installment, but his doctor then gave him the choice of lying with his feet higher than his head or an operation. Choosing the former required him to dictate much of the novel, though he was able to correct proofs. He eventually became “less and less dependent on dictation, writing the final sections of the manuscript in his own hand.” He mentions writing parts of installment twelve in a letter (Milgate 2006: 204), and this suggests he also wrote part thirteen.

To test for any dramatic effect of the switch to dictation, I divided the novel into four parts, the handwritten installmens1-4, the dictated 5-11, and the partially or wholly handwritten 12 and 13, and further divided these installments into sections of about 9,000 words. As can be seen in Fig. 1, the sections of the novel strongly tend to group together chronologically, though the beginning, 1-4 HW (1), is somewhat unusual, as often happens with the beginnings of novels, and 5-11 D (5) is also an outlier. At first, Fig. 1 seems to support a distinction between handwritten and dictated parts, but the presence of the largely handwritten installments 12 and 13 among the dictated parts suggests that narrative structure is a more potent force than mode of composition. Further work will be necessary to test other characteristics of the text, but the analysis shown in Fig. 1, along with many others based on shorter sections, does not suggest that Hardy's mode of composition radically affected his style.

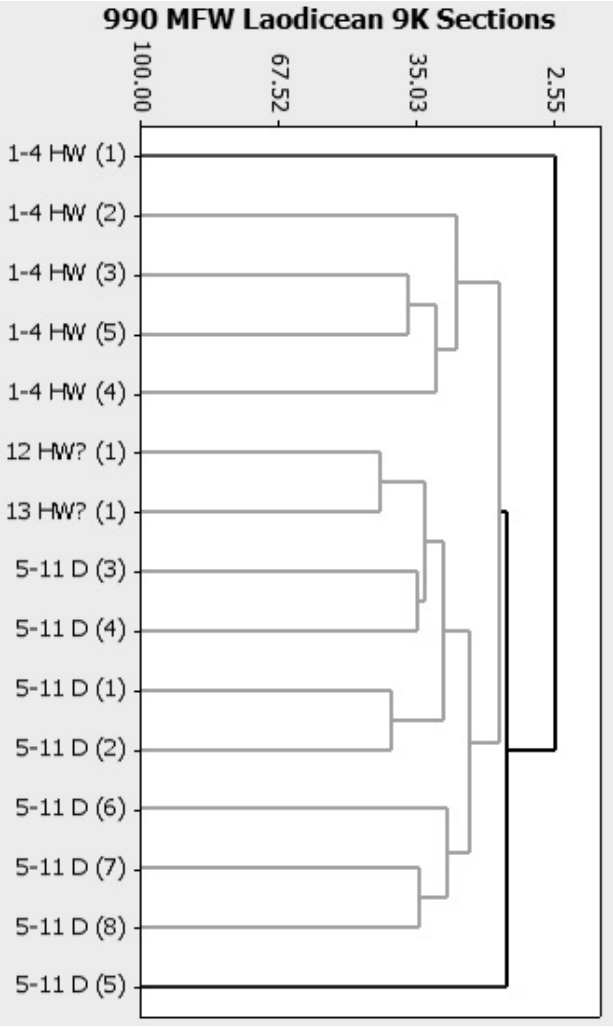


Fig. 1–Handwriting and Dictation in *A Laodicean*

Joseph Conrad presents a more complex problem. Several of his texts were partly dictated, including three I will examine here, the novellas *The End of the Tether* and *The Shadow-Line*, and novel *The Rescue*. Conrad dictated the second serial installment of “The End of the Tether” to Ford Maddox Ford under time pressure after part of the manuscript was accidentally burnt. I have separated the beginning and the burnt installment from the rest of the story, and have analyzed the parts in sections of about 2,600 words. As Fig. 2 shows, the first two sections of the dictated (burnt) installment

cluster with the handwritten beginning of the story, while the last section clusters with the handwritten rest of the story. The narrative structure is again quite clear here, but there is nothing to suggest that dictation altered Conrad's style.

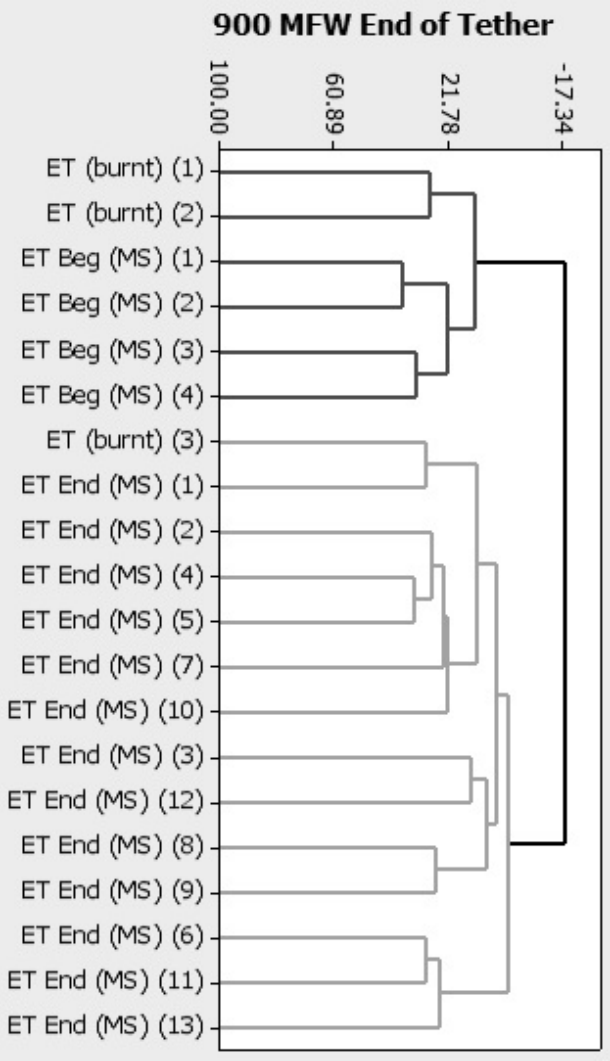


Fig. 2–Handwriting and Dictation in "The End of the Tether"

About one-fourth of *The Shadow Line*, beginning a little more than half-way through the novella, were dictated. Conrad himself ponders the possible effect of mode of composition, suggesting that "it will be curious for critics to compare my dictated to my written manner of expressing myself" (Conrad 1983: 543) . As with *The End of the Tether*, however, there is little evidence of any affect of dictation on the style of the novel, as Fig. 3 shows. Analyses based on different numbers of words vary somewhat, but the separate cluster containing the first four sections of the novel is very stable, and all analyses group the fifth handwritten section of the beginning of the novel with the dictated sections that immediately follow it. Again, narrative structure trumps any effect of the change in mode of composition.

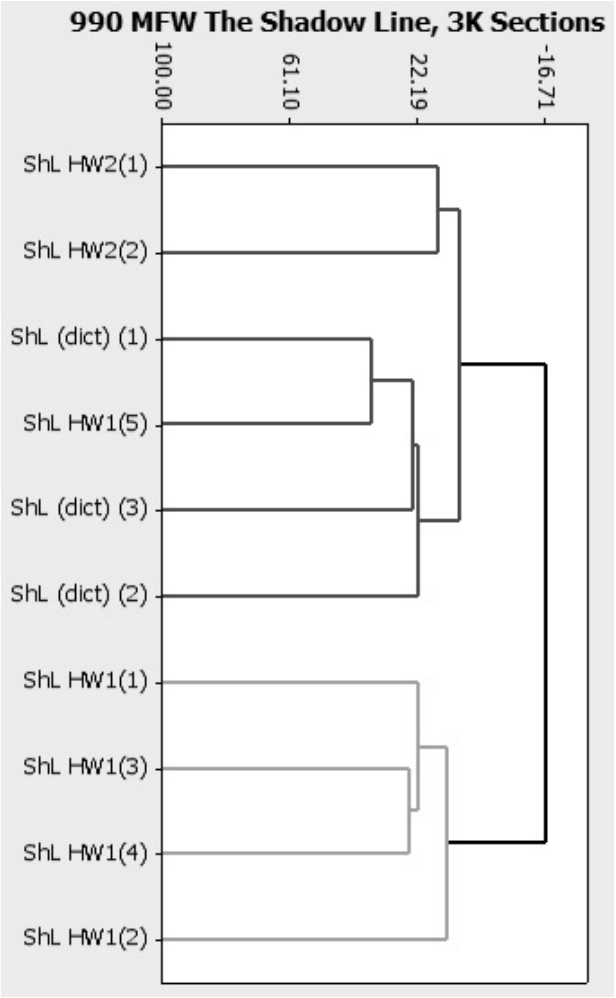


Fig. 3–Handwriting and Dictation in *The Shadow-Line*

The composition of *The Rescue* is unusual in that Conrad first worked on this novel from 1896 to 1898, but did not finish it until 1918-1919. The early part was handwritten, while the end was dictated. Although Conrad suggests that the novel might prove interesting as a case of style evolution, readers have found the style "homogeneous" (Karl 1979: 816). Here I divided the novel into dictated and handwritten parts, and cut them into sections of about 5,000 words for analysis. As Fig. 4 shows, the first two dictated parts cluster with the handwritten parts, and the 7th handwritten section clusters with the dictated parts. This pattern is extremely stable in analyses based on the 990-600 MFW, and all analyses show a mixing of sections produced by the two modes. Further tests based on other textual features will be needed to make the case more strongly, but, for Conrad, as for James and Hardy, the mode of composition has no obvious effect on style.

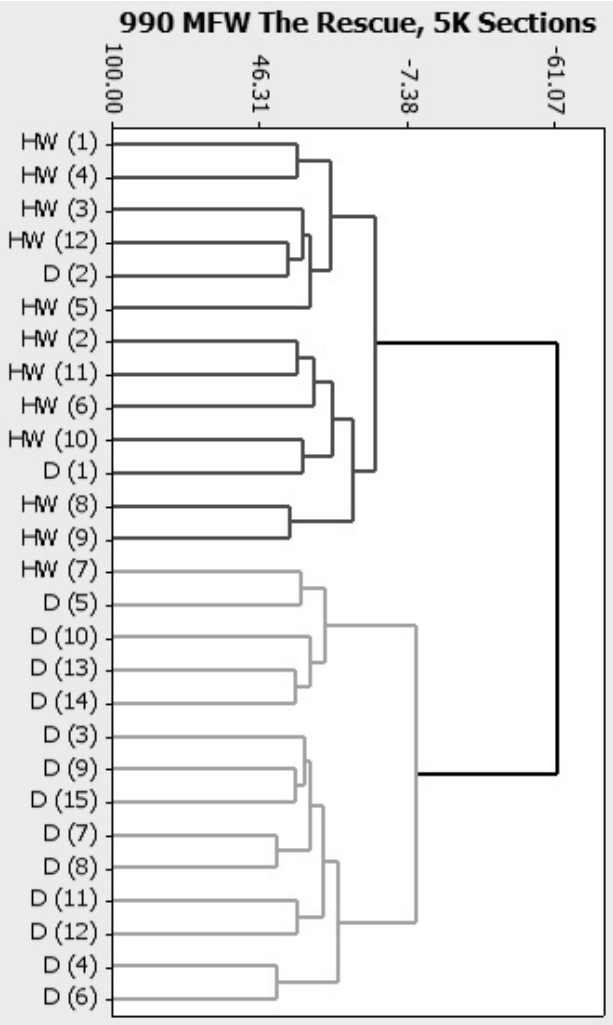


Fig. 4–Handwriting and Dictation in *The Rescue*

Turning to Walter Scott, we find a different scenario. About a third of the way through his career, Scott was writing very rapidly in an attempt to pay off an enormous debt. While writing *The Bride of Lammermoor* in 1818-19, he suffered from increasingly severe stomach pains (probably from gall-stone disease) that prevented him from writing, and finished the novel by dictation, though it is not entirely clear exactly where the dictation begins. The final extant MS leaf corresponds to Chapter 26 (of 33), but it ends with a catch word and has corrections for later leaves on the reverse (Milgate 1987: 170). We can be sure, however, that much, and probably most, of the last seven chapters were dictated. I have divided the novel into the part corresponding to the MS and the rest, and have divided both parts into sections of about 5,000 words. The important peculiarities of the pattern shown in Fig. 5 remain constant over many analyses. The first part of the novel for which no MS exists clusters (loosely) with the final two chapters of MS (though also with handwritten sections 6-12), but the handwritten sections 13-17 cluster with the handwritten beginning of the novel and the other three sections for which no MS exists. This peculiar pattern needs further study, but it does not support a difference between Scott's dictated and handwritten styles.

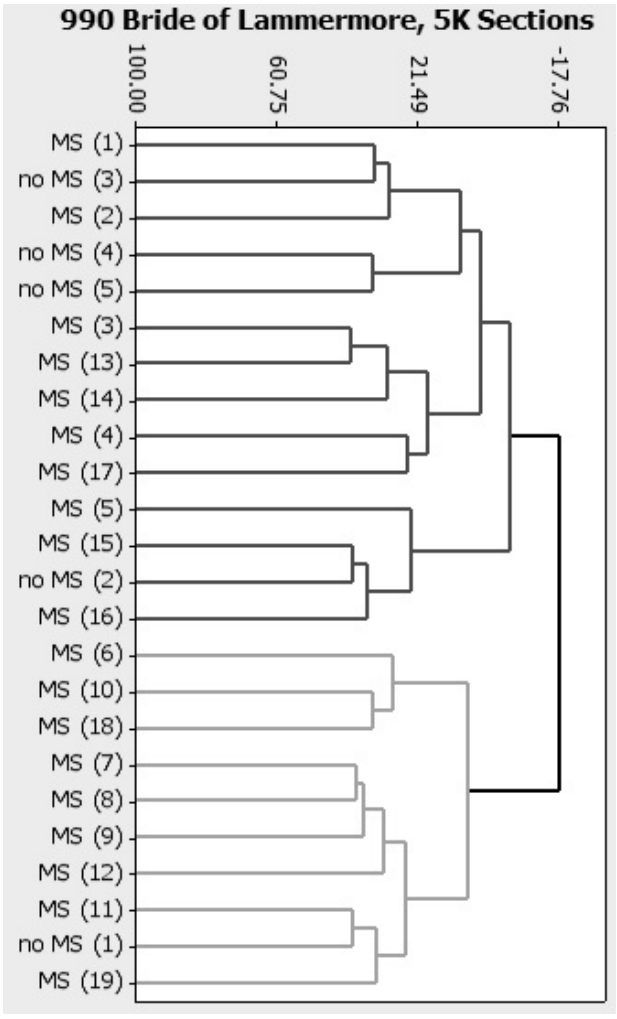


Fig. 5–Dictation and Handwriting in *The Bride of Lammermoor*

Because of continuing stomach pain, Scott also dictated about half of *Ivanhoe*, but finished the novel by hand after he began to recover. I have divided the novel into handwritten and dictated sections of about 9,000 words. The analysis in Fig. 6 shows that here, as before, there is no evidence of a significant shift in style when Scott's mode of composition changed. The last dictated section clusters with the following MS sections, while the next to last MS section, MS (9), clusters with the dictated sections. When fewer words are analyzed MS (9) shifts to the MS cluster, but Dict. (9) then also moves into the MS cluster. The most reasonable interpretation of this behavior is that the last dictated sections are similar to the adjacent

MS sections, and that the important factor is again narrative structure rather than mode of composition.

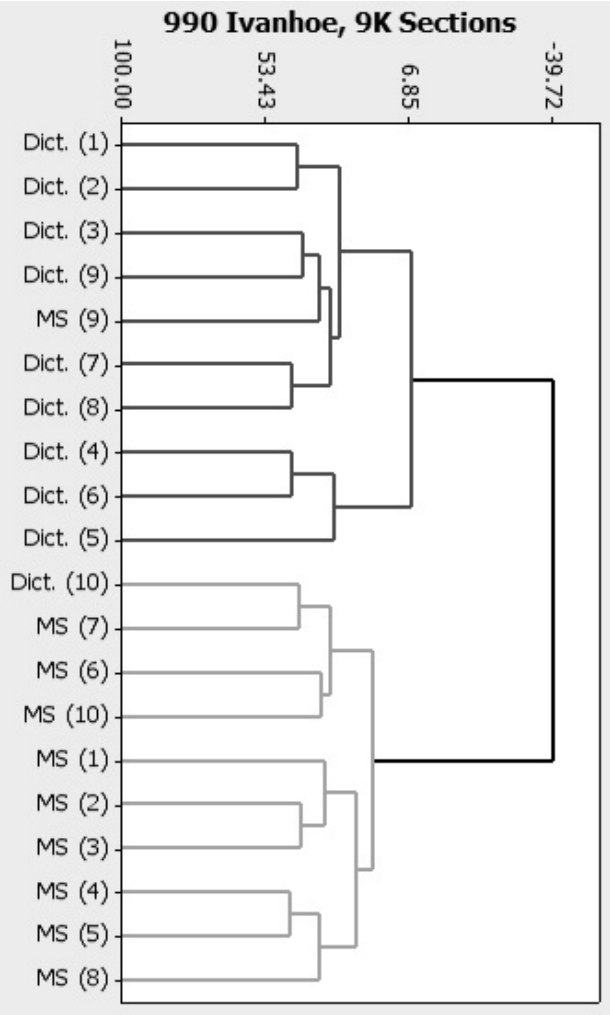


Fig. 6–Handwriting and Dictation in *Ivanhoe*

More analysis of other cases, especially those involving typewriting and word-processing, where some evidence for significant effects exists from composition studies, will be needed before any strong generalizations are possible. Yet the evidence from James and the three authors examined here strongly suggests that mode of composition has remarkably little effect on authorial style.[1]

References

Conrad, J. (1983). *The Collected Letters of Joseph Conrad.* F. Karl, L. Davies (eds.). Cambridge: Cambridge Univ. Press V. 5. .

Hoover, D. (2009). 'Modes of Composition in Henry James: Dictation, Style, and What Maisie Knew'. *Digital Humanities 2009.* University of Maryland, June 22-25, 2009.

Karl, F. (1979). *Joseph Conrad: The Three Lives, a Biography.* New York: Farrar, Straus and Giroux.

Millgate, Jane (1987). *Walter Scott: The Making of the Novelist.* Toronto: Univ. of Toronto Press.

Notes

1. The tempting conclusion that the authors' revisions may have erased any effects of the changes in mode of composition has some support from the heavy revisions of James and Conrad, but Hardy was apparently not a heavy reviser, and Scott famously revised very little.

Googling Ancient Places

Isaksen, Leif
leifuss@googlemail.com
Archaeology, University of Southamp

Barker, Elton
e.t.e.barker@open.ac.uk
Classical Studies, The Open University

Kansa, Eric C.
ekansa@ischool.berkeley.edu
School of Information, University of California, Berkeley

Byrne, Kate
k.byrne@ed.ac.uk
Informatics, University of Edinburgh

1. Overview

Our presentation about the Google Ancient Places (GAP) project will demonstrate new techniques to computationally identify places referenced in scholarly texts. We will also discuss deployment of simple Web services that use resulting place identifications to help bridge across online literary and material culture collections.

2. Project History

Funded through the Google Digital Humanities Award program (July 2010-June 2011), GAP <http://googleancientplaces.wordpress.com/> mines a portion of the Google Books Corpus <http://books.google.com/intl/en/googlebooks/history.html> to find books related to ancient locations identified by gazetteers of the Classical Mediterranean world.

GAP builds upon the Herodotus Encoded Space-Time Imaging Archive (HESTIA) project. HESTIA <http://www.open.ac.uk/arts/Hestia/> was a two-year collaboration (2008-2010) between The Open University and the Universities of Oxford and Birmingham, funded by the UK Arts and Humanities Research Council. Its aim was to explore new methods for visualizing relationships in Herodotus' Histories. The project explored multiple approaches, including:

1. mapping the frequency of references to specific locations (both spatially and in terms of linear narrative)
2. manually and automatically generating maps of the network connections between places.

The project made use of Greek and English versions of the text from the Perseus Digital Library <http://www.perseus.tufts.edu/> which are marked up with the Text Encoding Initiative (TEI) XML schema, including geographical locations based on automated string-matching with the Perseus internal gazetteer and Getty Thesaurus of Geographic Names. Closer analysis revealed that many of the locations were misidentifications, however, and a relatively labor-intensive process was required to correct them.

HESTIA's use of Perseus Digital Library resources demonstrates the growing power of open infrastructure already established in Classical studies. HESTIA also helped to demonstrate the utility of visualizing locations within a narrative. However, could the approach be automated so as to scale beyond manually processing individual texts? GAP attempts to answer this question through more sophisticated computational methods and by using additional open infrastructure, especially new Semantic gazetteers (see below) such as GeoNames <http://www.geonames.org/>, and the Pleiades Project <http://pleaides.stoa.org/>.

HESTIA's focus lies in a seminal text, the *Histories* by Herodotus. While primary and secondary literary sources represent key resources for Classical Studies, Classics also draws upon diverse sources of material evidence gathered from art history, architecture and archaeology (Mahony and Bodard 2010:3-5). These different sources of evidence and their associated scholarship are often highly "siloed". Reference to Perseus or Pleiades resources can improve their interoperability. To address this issue, GAP demonstrates how open digital humanities infrastructure together with the Google Books Corpus, can be used synergistically to bridge online literary and material culture collections. GAP uses Open Context <http://opencontext.org/> to test such services. Open Context is an open-access archaeological data publication system offering wide-ranging documentation of architecture, archaeological contexts, and objects from multiple contributors (Kansa and Kansa 2007). Open Context provides a map and timeline on its splash page to enable both providers and users to quickly identify related research. The ability to identify relevant scholarly literature relating to Open Context's material culture collections would be a ground-breaking extension to this service.

3. Approach

Prior experience with Herodotus' *Histories* informs GAP's methodology. In developing the HESTIA Narrative Timeline, we learned that places referenced in narrative texts generally cluster together to maintain narrative coherency. Given a set of toponyms with multiple possible identifications, the set of identifications with the shortest overall path between them is likely to be correct. In addition, we can add weight to the importance of each toponym by the number of possible locations it could refer to. Somewhat counter-intuitively, this means that small, obscure places with unusual names are much better guides to location than well-known places with many namesakes. While a useful starting point, several additional factors complicate accurate place identification:

1. The approach does not work well for fragments or with arbitrary higher-level structures such as the alphabetic organization of an encyclopedia.
2. The author may assume that the anticipated audience will be able to contextualize by other narrative elements (such as well-known individuals) and thus mention only a single location (or even none at all).
3. The author may contextualize by giving a territory in which the place is located. These can confuse point-based algorithms as there is no single 'best' point that represents them.
4. The author may have confused the place they are discussing with another, especially if they are commenting on another work or reporting independent sources.
5. Occasionally the pattern location clustering assumption simply does not hold. This is especially the case for places that do not perform an active function in the text such as personal names derived from places of origin (e.g. 'Herodotus of Halicarnassus').

Fortunately, new open infrastructure, especially Semantic Gazetteers such as GeoNames and Pleiades, can improve the precision of place identification. Both GeoNames and Pleiades offer open, machine-readable data curated by dedicated communities. They provide unique HTTP URIs for each place to which multiple names (toponyms), locations (such as spatial coordinates) and categories (like 'settlement' or 'region') can be assigned. These gazetteers make it much easier to handle the problem of synonymy and allow us to assign non-ambiguous and easily resolved public identifiers to places identified in the Google Books Corpus.

Nevertheless, even with the aid of gazetteers, the difficulties outlined above make place identifications highly probabilistic and uncertain, especially in cases where we find either insufficient or conflicting evidence. Hard cases can then be handled by a variety of methods, including more sophisticated but computationally expensive procedures or by manual effort of a scholar. Computationally, there are multiple levels at which we can look for clustering, including the chapter, book, and corpus (of the author or even genre). Looking at higher levels may provide us with broader contextual clues. A further advantage of working with massive digital corpora is that they frequently provide multiple translations and editions. In such cases we can use the linear chain of places in one edition to inform the processing of another and vice versa. Finally, as we process more books the system can record additional metadata about the places as well as the books. In particular it may see that in cases of homonymy, one location is much more frequently mentioned than all the others (such as the Egyptian Alexandria, as opposed to the many other cities of that name). This can help in cases where we have no other contextual clues to draw on. Google Books metadata and comparison of multiple editions found Google Books corpus may thus help resolve ambiguous place determinations in some cases.

It is also important to remember that there are some hard limits imposed on the process and some pragmatic aspects to our goals. First, we are only able to identify those places for which we have an entry in a gazetteer. Natural Language Processing available to us will not identify places previously unknown. Secondly, we are not looking for a 'perfect' set of results for the simple reason that natural language is ultimately indeterminate. Continued improvement of computational methods, as well as more traditional forms of scholarship will be required.

4. Outcomes

Scaling and adapting text processing methods developed for HESTIA for the larger Google Books Corpus represents one of the key challenges for GAP. To help evaluate the effectiveness of our approach, we first reconciled local identifiers used by the HESTIA project with Pleiades URIs. We will report on how our algorithmic approach to place identifications compares with places manually identified in HESTIA using the same raw text of the Histories as used by HESTIA. We will then report on results of our algorithmic method

on the 1828 translation of the Histories provided by Google. Finally we will discuss application of our algorithms for general use on the Google Books corpus, focusing on public domain texts with Library of Congress Headings DE-DG (Greco-Roman World; Greece; Italy).

GAP provides processing results in RDF-expressed annotations for each text. Such annotations are extremely useful to software but generally less helpful for humanities researchers who typically require a human interface. Thus, GAP also provides Web mapping tools, like those on the HESTIA and Open Context websites. These interfaces enable searches in both directions – from text to places, and from a place to the texts which reference it. To lower adoption barriers, we chose RESTful Web service (based on the Atom Syndication Format and GeoJSON) design patterns (see Blanke et al. 2009; Kansa and Bissell 2010). Such services enable other developers and digital humanists to incorporate our results into other research environments and applications.

As discussed above, the GAP project makes extensive use of existing digital humanities infrastructure, especially place gazetteers such as Pleiades. In doing so, GAP helps to demonstrate the growing maturity of digital scholarship in Classical studies. Rather than standing alone as isolated, one-off efforts, digital projects increasingly complement one-another and enable future work. In this light, we hope GAP will catalyze continued research (see Rosenzweig 2007) in the text processing methods, systems design, and semantic standards required to bridge gaps across literary and material culture collections.

References

Barker ETE, Bouzarovski S, Pelling CBR, Isaksen L. (2010). 'Mapping an ancient historian in a digital age: the Herodotus Encoded Space-Text-Image Archive (HESTIA)'. *Leeds International Classical Journal.* 9: 1-24.

Blanke, T., M. Hedges, and R. Palmer (2009). 'Restful services for the e-Humanities — web services that work for the e-Humanities ecosystem'. *Digital Ecosystems and Technologies, 2009. DEST '09. 3rd IEEE International Conference on Digital Ecosystems and Technologies..* Pp. 637-642.

Cohen, Daniel J. and Roy Rosenzweig (2006). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web.* Philadelphia: University of Pennsylvania Press. `http://chnm.gmu.edu/digitalhistory`.

Kansa, Eric C., and Ahrash N. Bissell (2010). 'Web Syndication Approaches for Sharing Primary Data in "Small Science" Domains.'. *Data Science Journal.* 9: 42-53.

Kansa, E., and S. Whitcher Kansa (2007). 'Open Context: Collaborative Data Publication to Bridge Field Research and Museum Collections'. *International Cultural Heritage Informatics Meeting (ICHIM07): Proceedings [J. Trant and D. Bearman (eds)].* Toronto: Archives & Museum Informatics. `http://www.archimuse.com/ichim07/papers/kansa/kansa.html`.

Mahony, Simon and Gabriel Bodard (2010). 'Introduction'. *Digital Research in the Study of Classical Antiquity [Simon Mahony and Gabriel Bodard (eds.)].* London: Ashgate, pp. 1-14.

Rosenzweig, Roy (2007). 'Collaboration and the cyberinfrastructure: Academic collaboration with museums and libraries in the digital era'. *First Monday.* 12(7).

Detecting and Characterizing National Style in the 19th Century Novel

Jockers, Matthew
mjockers@stanford.edu
Stanford University

In *Representative Irish Tales*, Yeats identified two basic categories of Irish fiction characterized by what he called "two different accents, the accent of the gentry and the less polished accent of the peasantry" (Yeats 1979). Writing of this distinction, John Cronin notes in The *Anglo-Irish Novel* how "Maria Edgeworth and William Carleton fit obviously enough the two extremes Yeats has defined but a middle-class figure like Gerald Griffin belongs a little uneasily somewhere in between" (Cronin 1980). Other critics including Thomas MacDonagh (MacDonagh 1916), Thomas Flannagan (Flanagan 1959), and most recently Charles Fanning (Fanning 2000) have all focused attention on the specific use of language in Irish narrative and the extent to which linguistic style and choice of theme and form reflects, or does not, the unique position of these Irish and Anglo-Irish writers in a country where the use of English was to evolve in a rather dramatic fashion.

Though Mark Hawthorne has written that the "Irish were not accustomed to the English language and were unaware of its subtleties and detonations" (Hawthorne 1975), Charles Fanning has argued that the Irish in fact became masters of the English language and employed a mode of "linguistic subversion" that allowed them to comment upon and even satirize the British who all the while seem to miss the point that the joke is on them (Fanning 2000). Cronin argues along similar lines to Fanning when he writes that the. . . idiomatic unease in their novels is not caused by any lack of ability on their part in the writing of a standard English idiom. It derives, rather, from the tangled situation in which they find themselves as novelists, directing their efforts towards an English-speaking public but trying to give that public a creative insight into a linguistically piebald area . . . they turned their very difficulties in regard to idiom to constructive account by confronting head-on the blending of the two idioms and two cultures. . . they turn this linguistic ragout to splendid account, making use in the process of English, Irish, and Anglo-Irish" (Cronin 1980).

The subject of this research paper, then, is the matter of exactly how 19th century Irish novelists uniquely employ style, setting, and theme. The critics seem to agree that something specific is going on in terms of language, form, and setting, and yet none gets to the heart of the matter, to the details of the prose and to the specific uses of language. Leveraging the tools and techniques from the authorship attribution and computational text analysis literature—specifically natural language processing, machine learning, and topic modeling—this paper compares and contrasts both linguistic style and narrative theme in a corpus of over 500 British and Irish novels from the 19th century. The results of this work show the precise extent to which Irish prose is stylistically different from English prose, and I identify and explore those linguistic and thematic features that mark the Irish novel as distinctly different from the British.

Specifically, my research examines style through an analysis of sentence and word level features. The results show, among other things, that Irish writers tend toward expressions that are both longer and more indeterminate than their British counterparts. Favoring the long sentence and greater use of the comma, the Irish write in comparatively complex, flowing sentences that favor (as measured by relative frequency) words denoting indeterminacy, words such as "most," "some," "may," and "yet." British writers, on the other hand, show a preference for shorter, more determinate sentences featuring words such as "know, never, no, nothing, must, not, only, all, should, last, first, and great." This result tends to confirm anecdotal observations made by scholars, including (Cronin 1980) who suggest that though the Irish may have sought to imitate and appeal to the stylistic preferences of a British dominated industry, they ultimately invented their own style of prose, which captured both the rhythms of the local language and the anxieties of a country struggling with its position vis-à-vis the colonizing presence of the British.

In addition to probing and comparing the stylistic habits of the two nations, this work further analyzes the prose at the level of theme and argues that there is an important link to be made between style and theme in Irish prose. To harvest latent themes, I employed the unsupervised topic modeling tools of the UMASS machine learning toolkit (McCallum 2002). A run of the model, which sought to identify the 25 most prominent topics in the corpus, resulted in one particular topic appearing with greater frequency in the Irish novels of the corpus. This topic, which was labeled as "the big house theme," is composed of words clearly relating to tenant-landlord relations and the familial issues

that are so often explored by Irish writers attempting to characterize these troubled relationships. The big house theme was found to be the most prominent topic in 35% of the Irish novels analyzed in this corpus, and it is present to a lesser degree in many of the others.

My analysis concludes by tracing the links between distinctly Irish themes and the elements of Irish style identified in the first part of the research. From the macroanalytic data derived at the corpus level, I present a chronological charting of Fanning's notion of linguistic subversion, and then I move to the micro level and offer a closer reading of several exemplary passages from works in the chronology. I discuss how linguistic subversion is inherent to the tradition of the "Irish Bull" and offer a brief discussion of Richard and Maria Edgeworth's 1835 essay on the subject in which they write with some humor that: "English is not the mother tongue of the natives of Ireland; to them it is a foreign language, and consequently, it is scarcely within the limits of probability, that they should avoid making blunders both in speaking and writing . . . Indeed, so perfectly persuaded are Englishmen of the truth of this proposition, that the moment an unfortunate Hibernian opens his lips they expect a bull, and listen with that well known look of sober contempt and smug self satisfaction, which sufficiently testifies their sense of safety and superiority." (Edgeworth 1835)

As early as *Castle Rackrent* (1800), Edgeworth had demonstrated her own command of linguistic subversion and an acute awareness of how to form her narrative and bend language to provide not simply a distinctly Irish novel but a seminal novel within the larger novelistic tradition. My work provides quantitative evidence of how, where, and why Irish style is different from British.

References

Cronin, J. (1980). *The Anglo-Irish Novel.* Totowa, N.J: Barnes & Noble Books.

Edgeworth, M. (1835). *Tales and Novels.* New York: Harper & brothers.

Fanning, C. (2000). *The Irish Voice in America : 250 Years of Irish-American Fiction..* Lexington, KY: University Press of Kentucky.

Flanagan, T. (1959). *The Irish Novelists, 1800-1850.* New York: Columbia University Press..

Hawthorne, M. D. (1975). *John and Michael Banim (The "O'hara Brothers") : A Study in the Early Development of the Anglo-Irish Novel, Romantic Reassessment..* Salzburg: Institut für Englische Sprache und Literatur, Universität Salzburg.

Macdonagh, T. (1916). *Literature in Ireland: Studies Irish and Anglo-Irish.* London: T. F. Unwin..

Andrew Kachites McCallum (2002). "Mallet: A Machine Learning for Language Toolkit.". `http://mallet.cs.umass.edu`.

Yeats, W. B. (1979). *Representative Irish Tales.* Atlantic Highlands, N.J.: Humanities Press.

Geo-Temporal Argumentation: The Roman Funeral Oration

Johanson, Christopher
cjohanson@gmail.com
UCLA Classics

1. Overview of the Discipline-Specific Project

The Roman aristocratic funeral of the Republic was an incredible show. It packaged the Roman spectacular trifecta, the procession, the eulogy and the subsequent games, which comprised gladiatorial and dramatic performances. While each of these components of the funeral has received individual treatment—in the case of the gladiatorial games, extensive—no detailed, comprehensive discussion of the aristocratic funeral of the Republic exists. Moreover, before gladiatorial games were held in the Colosseum and before dramatic performances were staged in a monumental theater, they were first held in ad hoc venues in the heart of Rome. No attempt has been made to situate the phenomenon within its surrounding context, the Roman Forum. My current digital/analog manuscript project, Spectacle in the Forum: the Roman Aristocratic Funeral of the Middle Republic, offers the first attempt to study the mid-Republican funeral in its totality and, in so doing, examines the most significant aspects of spectacular stagecraft of the Roman Republic.

2. The Intellectual Problem

Spectacle has received considerable attention in recent years, but its study has been marred by deficiencies in method. Classics scholar Richard Beacham pinpoints the problem: “Spectacle is three-dimensional and sequential, realized by taking place over a period of time, and its place, circumstance, and unfolding fundamentally shape what an audience both expects and experiences.” Ritual parades, political speeches, and religious rites are well described in ancient texts and frequently depicted in art. Yet, most spatial and spectacular analyses attempt to reconstruct the monuments, imagery, actors and audience, which are inherently kinetic and multi-dimensional (changing over space and time), by means of textual description and two-dimensional plans.

The impact of monumental structures on Roman performers and their audiences, what could and could not be seen during their performance, as well as the significance of *monumenta memoriae*, directly affected the shows when first performed, and the reading and interpretation of the records subsequently examined by scholars. Performance “stages” of the mid-Republic were ephemeral: extant temple podia, elevated balconies, and hillsides, might serve as caveae. Simple temporary structures may have been all that was needed to mount a production.

Three-dimensional digital models offer a partial solution. There are now a growing number of projects that have used computerized reconstructions to visualize Imperial Rome. There have been very few similar attempts to represent the Republican city, and hardly any that make scholarly arguments set within the digital reconstructions. Most reconstruction projects tend to focus on the creation of a highly accurate, extraordinarily precise digital model informed by scholarship as the ultimate goal. Instead, this project uses hypothetical reconstructions as a digital laboratory. By injecting historical context—the performers and the audience—into the digital environment, the digital investigation transforms the quantifiable elements of the ephemeral experience of ancient spectacle into a digital object fit for experiential analysis. It uses the hypothetical reconstructions as a digital laboratory to explore the staging of Roman spectacle and develop the digital toolset necessary for scholarly interrogation and publication of spatial and experiential arguments.

3. Geo-Temporal Argumentation: the Roman Funeral Oration

The *laudatio funebris* of the mid-Republic was genre-defying visual theater. While it is now generally agreed that the persuasive techniques of oratory comprised verbal (the content and delivery of the speech) and visual elements (gestures charged with meaning and explicit visual and topographic references), the degree to which the choreography of the funeral eulogy subordinated the words of the speech has not been fully examined. For much of the audience, the visuality of the event eclipsed the aural content. The *laudatio*, like the *pompa* before it, relied on a basic set of quasi-formulaic visual cues to communicate with the audience, or at least, to communicate some ideas to some of the audience. To call the laudatio a speech alone, and to classify it within the realm of oratory without qualification is to misunderstand much of the purpose and the choreography of the event. In this presentation, I will put the event in

its proper place: the Forum. Through the use of textual analysis, experiential investigation, and geo-temporal argumentation, I will demonstrate that the *laudatio funebris* was a multivariate theatrical event, comprising two discrete elements targeted at two distinct audiences.

Though one can use a laboratory built out of virtual world infrastructure to experiment, a researcher cannot (yet) "publish" the entirety of a laboratory experience and call it scholarly communication. Rather, the laboratory is the space where the research occurs; the results must be woven together into a narrative in order to engage with the larger scholarly conversation. Nonetheless, a text and image narrative is insufficient to convey the totality of the kinetic and temporal subject matter. Geo-temporal argumentation presents an innovative and more robust method of idea dissemination by offering:

1. Continuous and persistent spatial context.
2. Nodal points of departure for reader-based investigation
3. A refutable system embedded in a geographic context.
4. Narrative, perhaps *rhizomatic*, that enables non-linear review and exploration

When the experience and creation of kinetic transitions are fundamental to an understanding of an argument the reader must, quite simply, walk in the footsteps of the authors in order to participate in the debate, critique the result, and modify the conclusions. In this paper I aim to demonstrate that, for space- and time-centric, phenomenological investigations, geo-temporal argumentation is a new and superior form of scholarly communication.

4. The Technology and the Collaborative Project

As is always the case, digital humanities projects are collaborative endeavors. My "manuscript" project provides the domain-specific area of inquiry, but the digital platforms that facilitate the research are part of two, larger collaborative efforts of which I am but one of a number of co-investigators.

GeoTemporal Publication Platform: The research results and assessment will be published within HyperCities, a geo-temporal content aggregation and publication platform. Rather than create an entirely new digital humanities tool, "chapters" from my manuscript are being used as case-studies to guide the development of 3D narrative and mark-up tools within the HyperCities platform that will facilitate exploration of the data and publication of this new form of scholarly inquiry. We anticipate a mid-winter release of the working 3D system.

5. Project Samples

1. For the geo-aware 3D content, (Google Earth Required)
2. For a rough, sample narrative
3. An experiment in geo-temporal argumentation is now available in the inaugural issue of a new hybrid print/digital issue of the JSAH

The Object of Platform Studies: Relational Materialities and the Social Platform (the case of the Nintendo Wii)

Jones, Steven E.
s3jones1@gmail.com
Center for Textual Studies and Digital Humanities, Loyola University Chicago

Thiruvathukal, George K.
thiruvathukal@gmail.com
Center for Textual Studies and Digital Humanities, Loyola University Chicago

Racing the Beam: The Atari Video Computer System, by Ian Bogost and Nick Montfort, inaugurated the Platform Studies series at MIT Press in 2009. We've coauthored a new book in the series (currently under contract and final review, MIT Press), *Codename: Revolution: the Nintendo Wii Video Game Console*. Platform studies is a quintessentially Digital-Humanities approach, since it's explicitly focused on the interrelationship of computing and cultural expression. According to the series preface, the goal of platform studies is "to consider the lowest level of computing systems and to understand how these systems relate to culture and creativity." In practice, this involves paying close attention to specific hardware and software interactions--to the vertical relationships between a platform's multilayered materialities (Hayles; Kirschenbaum), from transistors to code to cultural reception. Any given act of platform-studies analysis may focus for example on the relationship between the chipset and the OS, or between the graphics processor and display parameters or game developers' designs. In computing terms, platform is an abstraction (Bogost and Montfort), a pragmatic frame placed around whatever hardware-and-software configuration is required in order to build or run certain specific applications (including creative works). The object of platform studies is thus a shifting series of possibility spaces, any number of dynamic thresholds between discrete levels of a system. As with the "text" in recent textual studies (McKenzie; McGann), the "platform" in platform studies is actually observed in action, as one or more transactional events, defined in the act of observation or performance. In this sense, platform studies examines the ways in which material conditions of computing systems determine (by constraining and affording) the experience of creative or expressive works.

Although there are now competing systems coming on to the market from Microsoft (Kinect) and Sony (Move), Nintendo's Wii (2006) was the first major video game console to be based on motion control, to make use of what Jesper Juul has called a mimetic interface to capture player movements in physical space and represent those movements in the game world. This paper will explain and demonstrate in precise terms how that's made possible through the use of a relatively off-the-shelf multichannel system of infrared (IR) and Bluetooth communications, along with data collected dynamically by the Analog Devices ADXL330 triple-axis accelerometer, a MEMS (Micro Electronic Mechanical Systems) device, a chip inside the wand-like Wii Remote--a tiny machine, with moving parts that measures motion along X, Y, and Z axes in relation to gravity. In this paper we'll show how this piece of hardware with accompanying code is used to design new software for the Wii, and we'll illustrate some of the specific constraints and affordances of the accelerometer by tracing the particular case study of third-party software developer Ubisoft's Samurai-Western FPS games, Red Steel I and II. The two games in this franchise mark points on a brief timeline, starting with the release of the first realistic action adventure game in 2006, as an exclusive title for the original Wii platform, followed by its relative critical failure, and then, four years later, the subsequent introduction of a "reboot" sequel, with completely revamped, cel-shaded, manga-style aesthetics, and all-new swordplay mechanics. This second game was timed to coincide with Nintendo's release of the controller add-on, Wii MotionPlus, a second MEMS device that plugs into the base of the Wii Remote and contains a "tuning-fork" gyroscope (the Invensense IDG-600), and thus adds more measurable dimensions to the system, giving the controller pseudo-global sensitivity and something closer to 1:1 mapping of player motions with in-game events. (This retrofit solution has now been superseded with the release of a new Wii Remote that ships with integrated versions of the two MEMS devices combined in one controller.)

That timeline, from 2006-2010, clearly marked a learning curve for Ubisoft's developers (including moving to the LyN 3D game engine, designed for the Wii, to make the second game), and the radical changes they made over the intervening four years serve as a vivid example of creative works being shaped in response to (and anticipation of) the constraints and affordances of a specific platform. The

story of the two *Red Steels* highlights an important fact--that a video game console is in effect a platform for production, transmission, publishing, and player reception, all in one system. In this way, a game platform--even at the lowest level--is an inherently social phenomenon, determined by developer as well as consumer and player expectations. Drawing from textual studies' idea of the social text, we'll argue that the "social platform" that is the object of platform studies is "social" in a particular sense: it's based on the relational materialities that both constrain and afford the production, transmission, and reception of creative works, the whole process that links hardware and software design to the player and the wider culture. This paper will illustrate this larger theoretical point by explaining one chain of concrete materialities on multiple levels in the case of Red Steel for the Wii--from tiny MEMS devices to in-game graphics, programming to marketing, transistors to cultural contexts.

References

Bogost, Ian and Nick Montfort (2009). *Racing the Beam: The Atari Video Computer System..* Cambridge, MA: MIT Press.

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

Jones, Steven E. and George K. Thiruvathukal (Forthcoming, 2012). *Codename: Revolution: The Nintendo Wii Video Game Console.* Cambridge, MA: MIT Press.

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

McGann, Jerome J. (1985). *Textual Criticism and Literary Interpretation.* Chicago: University of Chicago Press.

McKenzie, D. F. (1986). *Bibliography and the Sociology of Texts. The Panizzi Lectures, 1985..* London: The British Library.

The Time Machine: Capturing Worlds across Time in Texts

Juuso, Ilkka
ilkka.juuso@ee.oulu.fi
University of Oulu, Finland

Opas-Hänninen, Lisa Lena
lisa.lena.opas-hanninen@oulu.fi
University of Oulu, Finland

Johnson, Anthony
anthony.johnson@oulu.fi
University of Oulu, Finland

Seppänen, Tapio
tapio.seppanen@oulu.fi
University of Oulu, Finland

1. Introduction

This paper describes a number of ways in which a temporally-sensitive electronic dictionary resource, the Historical Thesaurus of the Oxford English Dictionary (2 vols; Oxford, 2009 [=HTOED]), may be employed in the automatic dating of words and entire texts. We investigate how the text captures time: most expressly, how the residue of the present (or the different 'presents' of language history) have managed to become trapped in the linguistic matrix of a narrative so that we sense, for instance, the difference between a period being represented and the narrator's temporal positionality, or even the gap between an author and his or her narrative stance. Through computer-assisted means we analyze the impact of later historical and linguistic events on the reporting of earlier events. To this end, we have developed an automatic system for retrieving dating information and a colour-coded browsing interface for searching and viewing the time-coded text, calling it the 'Time Machine'.

Novels capture worlds, but however disparate the materials that may go into them, something of the space-time in which they have been written remains as a residue. This, in part, is a function of language itself: the instabilities, changes and, above all, affordances at any one moment of that linguistic mesh that Lotman (1990) might have called the semiosphere. In part, too, it is a function of what, within Cultural Imagology, might more concretely be called the texture of the iconosphere (Johnson 2005, 2006): the distinctiveness given to the world at any particular moment by the

concatenation of signifying objects present within it. This is why it is an attractive idea to apply a tool such as a time-coded dictionary to novels written at time t purporting to convey events taking place at a time t-x. Within this frame of thought, the case of the historical novel is a particularly pronounced one. By definition, the genre tries to capture something of the iconosphere of a world that has passed us by (even though its semiosphere may remain that of a contemporary reader). And even in cases where the linguistic texture of the semiosphere is deliberately archaized – or localized by the use of dialect forms – the residue of the present remains. As a test case, we examine Diana Gabaldon's Cross Stitch (1991): a text which flaunts traditional temporal typologies by figuring a protagonist who crosses from the iconosphere of the mid-twentieth century to that of the seventeenth century and becomes trapped there: perceiving the past in a lexis and syntax which palpably belong to a different age.

Time has previously been explored within documents in several ways. Some work has concentrated on identifying expressions of time within text in an attempt to build models of the succession of events. This has been particularly fruitful in the case of, for example, medical discharge records and road accident reports, where the sequence of events is of great importance (Hirschman 1981, Kayser and Nouioua 2009). Other work has used a training set of time-associated words and a Naïve Bayes Classifier to detect temporal concepts in blogs (Noro et al. 2006). While this work is promising in analyzing writings about daily life in a compact time frame, it seems ill-equipped for investigating iconospheres that deal with spans constituting years or even decades. Thus a tool that can retrieve time-related information from the HTOED automatically offers a very promising way forward for the literary and linguistic scholar.

2. Methodology

Using the 'Time Machine', we map out the iconospheric precision with which Gabaldon represents different characters in her fiction (not to mention the humour generated by the gradual blending of their discourses as the novel progresses). But beyond this, by linking our tool with the powerful additional resources which the HTOED has now opened up for those studying the 'external', 'mental' and 'social' worlds of the novel from a historical and etymological perspective, the project hopes to facilitate the achievement of a more nuanced understanding of the interrelationship between 'real' and 'fictional' time in the historical novel than has been possible before.

In order to better study iconospheres, we sought to develop a tool that would automatically look up dating information and definitions for words, processing entire texts at a time, thereby removing the need for manual queries using a dictionary. Furthermore, we wanted the tool, on the one hand, to enable users to specify time periods of interest for closer inspection while, on the other hand, it left them free to browse the material through diverse visualization schemes in order to discover trends or new time periods of interest.

At present the tool is a prototype, running inside a web browser, in order to enable rapid experimentation with new visualization schemes using CSS. We use a local SQL database to store the HTOED data. Texts can be uploaded via a browser interface and are processed in any user-defined units tagged in the text, e.g. page by page or speaker by speaker, or in the text as a whole. The tool reads both XML and plain text. To finish verifying that the visualization schemes we have chosen are useful, we wish to bring the tool to the digital humanities community, in addition to the poetics and linguistics community (Johnson et al. 2010). Following this, we intend to develop the final tool in Java for inclusion within the LICHEN toolbox.

3. Results

What the HTOED is able to offer to the 'Time Machine' is the ability to isolate different experiential modes within particular iconospheres at the same time as it reveals the range of etymological meanings open to the reader at any given moment. (This, of course, is an invaluable aid for critics who wish to avoid anachronism in their own readings.) In our preliminary development of the 'Time Machine' we concentrated on its capacity for isolating different lexical categories within a given iconosphere and indicating the etymological choices available for particular readings. At the top of the screen, the Source section allows the user to choose either an entire text or some part of it (see Figure 1 below: a case in which the speaker Jamie has been chosen). In the Filter section, the user can choose to narrow the search down to one particular word, or to all words that were in use at a particular time (choosing either first use date, last use date or both). To produce the present screen we started by choosing all the words that entered the language after 1742, in other words after the time period in which Jamie speaks. The Colour-coded text section then highlighted all those words which entered the language after the given date, as did our Wordlist section.

Our initial investigation found that the tool is able to pick out swathes of temporal incongruities from this playful text or, further, search out instances relating more specifically to the 'external', 'mental' and 'social' worlds of the novel. It spots moments when the eighteenth-century clansman Jamie seems prescient (mentioning 'aesthetics' for instance, or re-circulating the word 'sadist', which has been bandied to him by his twentieth-century wife). It detects instabilities not only in the iconosphere of the 1700s – which Gabaldon has carefully researched – but also in the representation of mid-twentieth-century England (which she appears to have taken more for granted).

However, despite the manifest advantage of using even this approach to the 'Time Machine' to spot faultlines and incongruities within the fictional world of a novel, some teething-troubles remained: the most significant being that, unlike human readers, the prototype cannot, of course, intuit the 'correct' lexical choice from the range of possible meanings thrown up by a search. Accordingly, we have tweaked the search and display capability of the 'Time Machine' so that it can also narrow its lexical catchment area by trawling parts of speech (such as substantives) in which cultural and temporal change exhibit their highest visibility. Figure 1 demonstrates how, using these restrictions, the prototype is able to flag up the way in which Gabaldon has inadvertently endowed Jamie's lexicon with three words stereotypically associated with Scottishness ('Sassenach, shinty, sporran') which were not, in fact, recorded until some time after the period in which Jamie is meant to be speaking.

In sum, our study indicates that automated access to chronological information, such as the date of first use for any given word, and full etymologies has promising applications in literary and historical research that has until now relied mostly on intuition and laborious manual methods to combine dating information and texts. And beyond this, with some adaptation, it is also clear that the 'Time Machine' could be of significance within areas such as forensic linguistics, collocation studies, and the study of micro-linguistic change over time in large corpora.

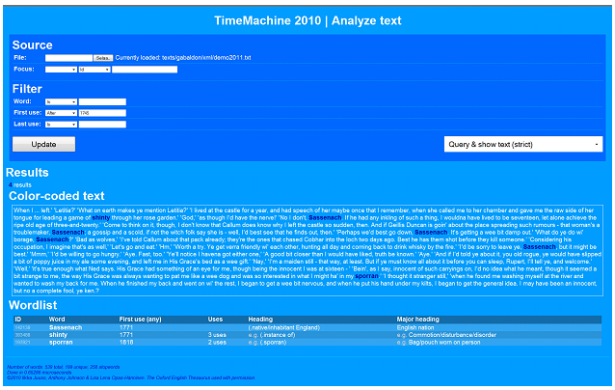


Figure 1

References

Gabaldon, D. (1991). *Cross Stitch.* Arrow Books Ltd..

(2009). *Historical Thesaurus of the Oxford English Dictionary.* Oxford: OUP.

Hirschman, L. (1981). 'Retrieving time information from natural-language texts'. *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*.

Johnson, Anthony W. (2005). 'Notes Towards a New Imagology'. *European English Messenger.* 1: 50-58.

Johnson, Anthony W. (2006). 'New Methodologies: Imagology, Language and English Philology'. *Linguistic Topics and Language Teaching.* Antilla, H., Gear, J., Heikkinen, A., Sallinen, Riitta (eds.) Oulu University Press7-27.

Johnson, Anthony W., Opas-Hänninen, L.L., Juuso, I. (2010). 'Stitches in Time and Switches in Text: Diana Gabaldon and the Historical Thesaurus of the Oxford English Dictionary'. *Paper presented at PALA2010.* .

Kayser, D., Nouioua, F. (2009). 'From the textual description of an accident to its causes'. *Journal of Artificial Intelligence.* 173: 12-13.

Lotman, Yuri M. (1990). *Universe of the Mind: A Semiotic Theory of Culture.* Ann Shukman (ed.). 173: 12-13.

Noro, T., Inui, T., Takamura, H., Okumura, M. (2006). 'Time period identification of events in text'. *Proceedings of the 21st international Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (Sydney, Australia, July 17 - 18, 2006). Association for Computational Linguistics.* `http://dx.doi.org/10.3115/1220175.1220320`.

Trends 21 Corpus: A Large Annotated Korean Newspaper Corpus for Linguistic and Cultural Studies

Kim, Heunggyu
gardener@korea.ac.kr
Department of Korean Language and Literature, Korea University

Kang, Beom-mo
bmkang@korea.ac.kr
Department of Linguistics, Korea University

Lee, Do-Gil
motdg@korea.ac.kr
Research Institute of Korean Studies, Korea University

Chung, Eugene
echung2@korea.ac.kr
Research Institute of Korean Studies, Korea University

Kim, Ilhwan
ilhwan52@gmail.com
Research Institute of Korean Studies, Korea University

1. Introduction

This study aims to introduce how a Korean newspaper corpus, Trends 21 has been constructed and to explore how social, cultural, and linguistic characteristics are portrayed in the Trends 21 corpus. Newspapers contain enormous quantities of language resources which mirror social and cultural characteristics as they undergo gradual as well as sudden changes. Newspapers are regularly published and contain stories of events, personalities, crimes, business, entertainment, society, sports and others. Editorials discuss current or recent news of either general interest or a specific topic. Journalists are trained to write objectively and to show all sides to an issue. In addition, the sources for the news story are identified and are reliable. Therefore, we have employed the newspaper corpus to identify social or culture trends.

2. ***Trends 21*** Project

2.1. Aims and Background

Trends 21 is the name of a project within the government-led humanities promotion program.[1] The Research Institute of Korean Studies at Korea University has developed the *Trends 21* corpus, a collection of four major Korean daily and national newspapers issued from the year 2000. The goal of the *Trends 21* project can be summarized with the following three points: first, to construct language resources of newspaper articles as a large general purpose database; second, to identify linguistic/social/cultural characteristics and to analyze their changes in Korea; finally, to measure and to estimate any linguistic/social/cultural trends from patterns of language use. One of the outcomes from this project is the *Trends 21* corpus. It is a collection of Korean newspaper texts covering most of the topics in print. In the next section, we present how the *Trends 21* corpus has been built.

2.2. Designing and Compiling the Corpus

In order to achieve the project goals, articles were culled from a number of newspaper companies. We collected newspaper articles from four major daily national newspapers issued in Korea for one decade, between 2000 and 2009.[2] The candidate dailies are Chosun, Dong-a, Joongang, and Hankyoreh.

These daily newspaper companies have provided us with all the contents printed for ten years. In electronic form, the majority of newspaper services are provided in various Standard Generalized Markup Language (SGML) format, which is hard for us to unify the format by using NewsML. Due to this situation, we developed a 'Trends 21 Markup Language (T21ML)' in order to construct our raw corpus. With the availability of machine-readable texts, especially the collection of a large quantity of articles, it was possible to build a large-scale raw corpus. However, we did not upload all the contents from the newspapers into our corpus. Instead we eliminated irrelevant contents (like obituaries) in order to balance the contents.

For our research purposes we established twelve classes of content, namely 'T21 Class', to classify the contents of news articles, namely: politics, international news, economics, society, culture, sports, science, columns, opinions, special issues, regions, and people. It excludes lists of names, lists of stocks, obituaries, advertisements, and weather. Although

some contents are removed by design, our corpus contains various contents or topics as a whole. Saturation (McEnery *et al.* 2006) at the lexical level can be tested for representativeness of a corpus.

Once the raw corpus was constructed, we employed an automatic morphological analyzer and tagger for Korean, KMAT (Lee & Rim 2009), to annotate parts-of-speech and morpheme information. We applied two-stage tagging processes to our raw corpus, in which an available annotated corpus consisting of 15 million words is corrected by humans and then is employed as a training corpus for the tagger. Further, human annotators not only corrected the erroneous analyses produced by the tagging system, but also improved the tagging system by finding problematic fixed expressions, picking out homonyms, and classifying unseen types of borrowed words or proper nouns. During the first three-year phase of the *Trends 21* Project (2008-2010), this corpus has been fully annotated for parts-of-speech and morphological information. Figure 1 shows the processing architecture of building the *Trends 21* Corpus.

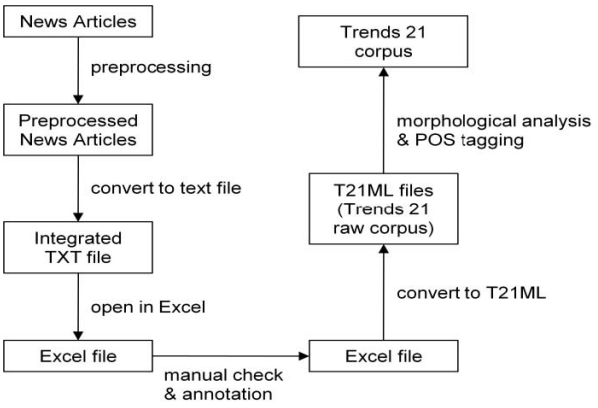


Figure 1. An overview of the *Trends 21*corpus building process

As of Oct 2010, the *Trends 21* corpus consists of about 400 million words, and it is by far the greatest morphologically annotated corpus of Korean. In Table 1, statistical information is provided.[3]

3. Case Study: Co-occurrence Network Analysis

In a case study, we focus on only the nouns that are included in the *Trends 21* corpus. A network based approach is then introduced that can deal with visualizing related nouns. According to Stubbs (1996), frequently occurring patterns allow the observer to make deductions about what a group or society sees as valuable or important. Information about collocation means that new concepts and the range of associations of a word can be monitored. We select target words and extract their co-occurring words appearing nearby. Co-occurrence analysis assumes that two semantically related terms co-occur in the same text segments (Sinclair 1991). In contrast to most previous studies that observe co-occurrences within the same sentence, we propose as a search window size a paragraph (Kang 2010). A paragraph of news article is highly coherent in that its sentences are related to one another to describe one short story or an event.

The extraction of co-occurring words is based on the statistical information about the co-occurrences of words. The mutual information or z-score has mainly been used in various studies as a statistical measure; however, both of the measures give skewed results to infrequently used words. To reduce this difficulty, we adopt t-score as a measure of how strongly word pairs (a target word and co-occurring words) are related (Kang 2010).

Then the information is represented as a network, a formal graph based approach. We have employed *Pajek* (Nooy, Mrvar & Vladimir 2005) for analysis and visualization of co-occurrence networks. The network structure typically consists of nodes connected by weighted links. Given the current data set, target words and co-occurrences assign a term or a concept to each node and the values of the t-scores to link. This network provides a graphic visualization of potential relationships between nouns that portray social/cultural trends with respect to their language use patterns.

Figure 2 is the co-occurrence network of thirty Korean emotional nouns, such as: ‘love’, ‘hatred’, ‘hope’, ‘disappointment’, ‘happiness’, ‘unhappiness’ and so on. In Figure 2, ‘father’ (*abeci* in Korean) co-occurs with ‘love’, ‘hope’, ‘happiness’, ‘hatred’, and ‘unhappiness’; on the other hand ‘daddy’ (*appa* in Korean) only co-occurs with ‘happiness’.

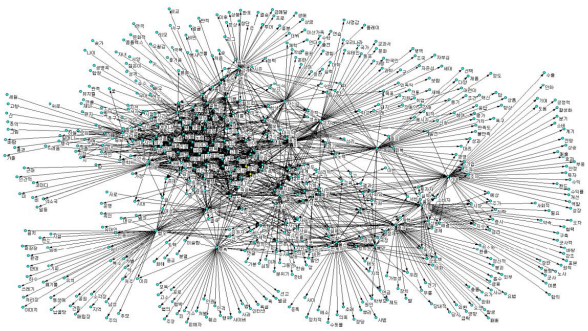


Figure 2. Network of thirty Korean emotional nouns with their fifty co-occurring words

The word ‘hatred’ co-exists with ‘terror’, ‘Islam’, ‘(human) race’, ‘religion’, and ‘media’. We notice that in the early twenty-first century there were many international conflicts. If we expand the number of co-occurrences, we may deduce different interpretations from the articles.

4. Conclusion

This paper has presented how the *Trends 21* corpus is built and how it is composed. We have proposed a visualization method to express co-occurrences of words in an overview network. The network approach to words in news articles represents contemporary Korean language use. Moreover, information about co-occurrences helps us understand social/cultural issues at a point of time.

The construction of the *Trends 21* corpus is not done yet. The same composition schema is going to be followed year by year in order for the corpus to be constantly updated. In that sense, the *Trends 21* corpus serves us as a monitor corpus (Sinclair 1991). This corpus can also reflect language changes in constant growth. In the future we would like to apply cluster analysis as well as keyword analysis. We further plan to enhance the network analysis by displaying concept hierarchy. Finally we also plan to investigate networks according to topics and co-occurrences within an article, not only with in a paragraph.

5. Table 1: Statistical Information of the ***Trends 21*** Corpus

Target	Unit	Size
Trends 21 Corpus	*ejels*[4]	348,261,978
Trends 21 Corpus	article	1,763,581
Trends 21 Corpus	paragraph	13,440,141
Common Nouns in Trends 21 Corpus	type	487,385
Common Nouns in Trends 21 Corpus	token	223,794,143

References

Biber, D., Conrad, S., Reppen, R. (1998). *Corpus Linguistics.* Cambridge: Cambridge University Press, .

Church, KW, Gale, W., Hanks, P, Hindle, D (1991). 'Using statistics in lexical analysis'. *Lexical Acquisition: Using On-line Resources to Build a Lexicon.* Lawrence Erlbaum, pp. 115-164, .

Kang, B. (2010). 'Constructing Networks of Related Concepts Based on Co-occurring Nouns'. *Korean Semantics.* V. 32, pp. 1-28.

Kim, I., Lee, D, Kang, B (2010). 'A Study of Emotion Nouns Based on Co-occurrence Relation Networks'. *Korean Linguistics.* .

Lee, D., Rim, H. (2009). 'Probabilistic Modeling of Korean Morphology''. *IEEE Transactions on Audio, Speech, and Language Processing.* 5: 945-955.

McEnery, T, Xiao, R., Tono, Y. (2006). *Corpus-Based Language Studies.* Abingdon: Routledge.

Nooy, W., Mrvar, A., Vladimir, B. (2005). *Exploratory Social Network Analysis with Pajek.* Cambridge University Press.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Stubbs, M. (1996). *Text and Corpus Analysis.* Oxford: Blackwell.

Notes

1. The Humanities Korea (HK) project is an initiative aiming to foster world-class research centers that carry out interdisciplinary studies in the areas of humanities.
2. A daily newspaper is issued every day with the exception of Sundays and some national holidays. A national newspaper, in contrast with a local newspaper that serves a city or region, circulates throughout the whole country.
3. This information is based on the compiled data between 2000 and 2008.
4. An ejel refers to a chunk between spaces in Korean. The ejel may be one word itself or the morphosyntactic combination of either one word and particle(s) or one word and ending(s).

Abstract Values in the 19th Century British Novel: Decline and Transformation of a Semantic Field

Le-Khac, Long
llekhac@stanford.edu
Stanford University

Heuser, Ryan
heuser@stanford.edu
Stanford University

1. Introduction

This paper analyzes the historical behavior of several semantic fields of "abstract values" in a corpus of 2,779 19th century British novels. The corpus is a composite archive of canonical and non-canonical texts drawn from Project Gutenberg, Internet Archive, and Chadwyck-Healey's 19th-century Fiction Collection. In his classic study, *Culture and Society*, Raymond Williams claimed that a group of keywords that arose and/or changed dramatically in the nineteenth century offered "a special kind of map [of the] wider changes in life and thought" of the age (Williams 1958). We develop Williams's insight by applying quantitative methods to a much larger corpus than available at the time of his study. Using a tool we built specifically for our research, we were able to aggregate words whose historical frequencies follow similar trends, thus identifying particularly dynamic semantic cohorts; from these, we found dramatic declines and transformations in fields of social restraint, moral valuation, sentiment, and partiality over the nineteenth century – and an equally dramatic increase in the use of concrete description fields in the same period. We examine the implications of these findings with respect to broader ideological and narrative patterns of the British novel.

2. Background

In prior applications of semantic fields to quantitative literary studies, researchers have tended to measure the relative presence of certain fields, "themes" or otherwise-labeled word-groups in individual texts (e.g. Louwese 2004; Ide 1989; Fortier 1989). We hope to complement such comparative work by tracing the diachronic behavior of particular semantic fields across a corpus of nineteenth-century novels. In addition, we aim to specify our theoretical object of the semantic field more precisely, both by developing our fields through an empirical method of word-cohort correlation, and by grounding them in their original conceptualization by early twentieth-century semantics. In "Bedeuntungssysteme," R. N. Meyer influentially defined a semantic field as "the ordering of a definite number of expressions from a particular point of view"—or in other words, from a particular "differentiating factor" (Meyer 1910). To borrow Meyer's example: the sense of purposefulness, present in the transitive verb *ersteigen* (to climb) but not in the intransitive verb *steigen* (to rise), could serve as the "differentiating factor" around which a particular *Bedeuntungssystem* derived its identity.

In its period focus and objectives, this project is indebted to Raymond Williams's *Culture and Society*, which analyzes the historical semantics of a period of unprecedented change for Britain. He contends that changes in discourse help reveal broader sociocultural changes. These wider changes, he argues, are of no small consequence; indeed, they introduced many of the social elements and ideas central to what we now think of as distinctive to our modern way of life (Williams 1958). Of course, Williams's ambitious attempt to analyze an entire social discourse, astonishing as it is, lacked the tools and corpora now available to digital humanities scholars. This paper represents some first steps in pursuing Williams's objectives by applying quantitative methods to a large novelistic corpus in order to explore specific but dramatic changes in language and culture in this volatile period.

3. Methods

Our method of field-creation consists of two stages: discovery and development. First, to *discover* potential fields, we developed a technique of word-cohort correlation. We input "seed" words considered significant by previous literary historical work and query the corpus for words whose historical frequency-trends most resemble those of the "seed" words. When this automatically generated cohort of correlated words shares a specifiable differentiating factor and their overall trend is significant, we consider such a word-cohort the embryo of a dynamic semantic field ripe for development. We then *develop* the field further by employing semantic taxonomies from within the humanities and linguistics such as the OED's historical thesaurus to identify the semantic content of these word cohorts and subdivide them into specific semantic fields to track. This method, which oscillates

between semantic taxonomies and empirical word frequency correlations, ensures that the semantic fields we generate satisfy two characteristics: semantic coherence and coherence of historical behavior. We consider this dual requirement as a pragmatic move. The first requirement ensures that our results are semantically and culturally interpretable. The second requirement ensures that the aggregate term frequency results of the semantic fields are actually representative of the behaviors of their constituent words.

4. The Semantic Fields

In particular, this paper reports on a study of the historical behavior of four semantic fields of "abstract values." We identified these transforming semantic fields through the methods described above. This study presents and analyzes the dramatic decline and transformation of four semantic fields [Fig. 1] discovered under this method (each is named after what we considered the differentiating factor of that field).

Field Name	Example Words
Values of Social Restraint	modesty, sensibility, propriety
Moral Valuation	virtue, sin, conduct
Sentiment	passion, sentiment, sensibility
Partiality	partiality, prejudice, disinterested

Fig. 1

5. Results

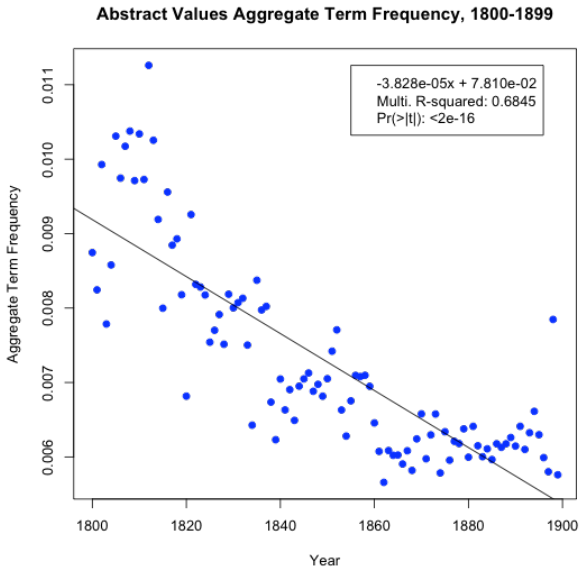


Fig. 2

Tracing the diachronic behavior of these fields over the nineteenth century, we found the four abstract values fields exhibited strikingly parallel downward trends [For their individual plots, see Fig. 5-8]. Collectively, the aggregate term frequency for the fields of abstract values decreases step-wise through the nineteenth century [Fig. 2], from ~1% of all words in the period of 1800-1810, to ~0.6% of words (~1 in every 170 words) by the 1860s, a decrease of about 40%.

6. Discussion

Given the range and scope of our corpus and the magnitude of this trend, we consider our results reflective of broad changes in the 19th century British novel. The data indicates a significant decrease in the usage of these fields. Without positing a simple reflective relationship between literary and sociocultural currents, we nevertheless take seriously Raymond Williams's approach to social changes through changes in discourse. Thus, we consider the data to suggest the responsiveness of novelistic language to fundamental shifts in British value systems and social norms in this turbulent and transformative period, specifically shifts away from values of restraint, virtue, objectivity, and sentiment.

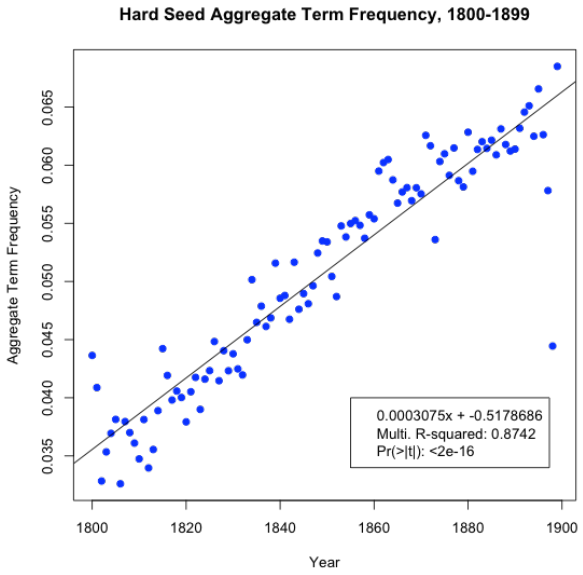


Fig. 3

The historical behavior data of an entirely different set of words, discovered under the same method, helps to contextualize and interpret this trend. Instead of a semantic field tightly organized around a specific differentiating factor, this highly correlated word-cohort (named "Hard Seed" after its seed word) comprises a variety of semantic fields and types of words—colors, body parts, numbers, locational and directional adjectives and prepositions, action verbs, and physical adjectives. This word cohort can be collectively characterized as concrete description words. In contrast to the values fields, the aggregate term frequency of this latter group [Fig. 3] increases steadily across the 19th century from 3.5% of all words (~1 in every 30 words) to 6.5% of all words (~1 in every 15 words)—an increase in usage of about 85%.

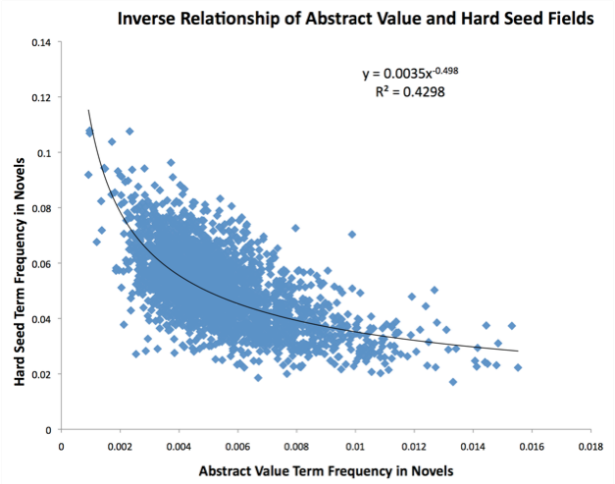


Fig. 4

Plotting the term frequencies of the abstract values field against those of the hard seed field in each of the novels in our corpus reveals a strongly inverse relationship between the two [Fig. 4]. Given the observed tendency for novels with higher frequencies of hard seed words to have lower frequencies of abstract values words and vice versa, we produced two rankings of the novels to see the types of narrative that correspond to the emphasis of one field over the other [see Fig. 9 for a ranking of a subset of the corpus, the Chadwyck-Healey fiction collection]. Strikingly, ranking novels by these two features indeed separates out clusters of genres into a spectrum. The resulting distribution of novels allows us to interpret these two major correlated historical trends in novelistic language as deeper shifts in narrative mode. The spectrum shows an overarching movement from narratives with small social spaces organized by highly polarized, evaluative, and uniform fields of social norms to narratives with far larger social spaces where the fields of social norms are more diverse, conflicting, ambiguous, and ultimately, less constraining. Simultaneously, there is a stylistic move away from abstract, explicitly evaluative language to concrete, physical language whose valuation, if any, is more ambiguous, variable, and indirect. That the expansion of narrative social space corresponds to this stylistic shift suggests a systemic tendency in which the representation of wider, more diverse, and less constraining social spaces is made possible by a more physical, concrete language.

7. Conclusion

This study represents initial steps in developing the quantitative analysis of the historical semantics of literary discourse in a corpus robust enough to allow the study of large-scale historical change. The methods developed herein have proven promising in identifying and analyzing robust semantic fields whose dynamics can rigorously be interpreted as reflections of literary and cultural trends. A wide field of potential inquiry remains for future studies in this vein.

8. Figures 5-9

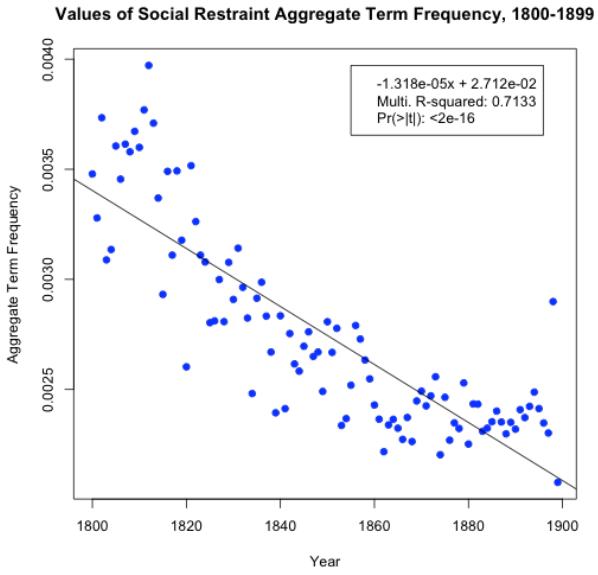


Fig. 5

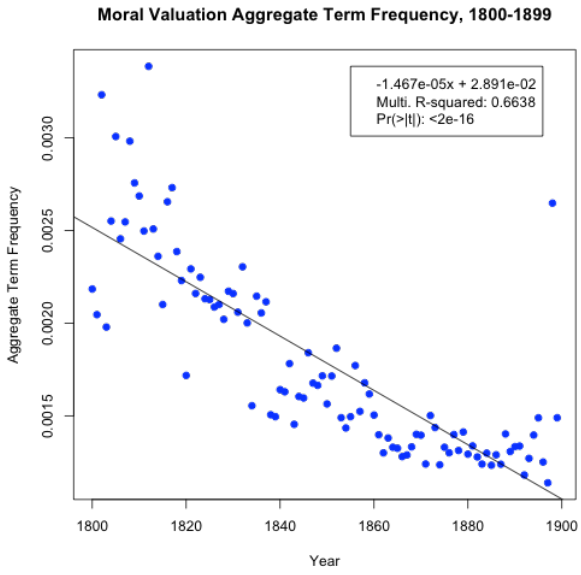


Fig. 6

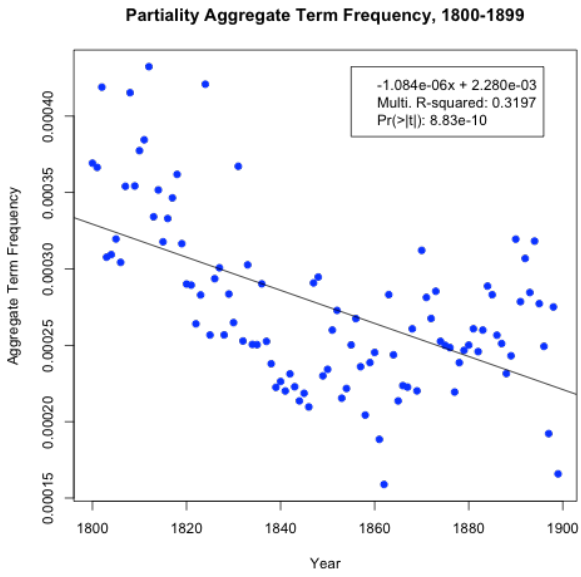


Fig. 7

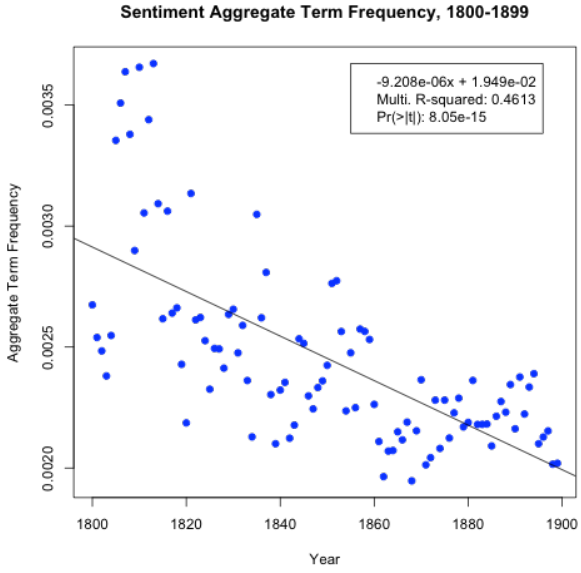


Fig. 8

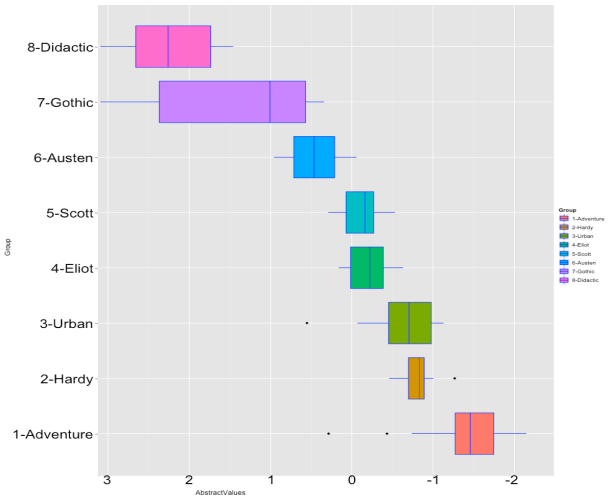


Fig. 9: Distribution of genres and authors from those with the most frequent usage of Abstract Values, to the least frequent. The X-axis indicates the number of standard deviations above or below the mean. The vertical line in the center of each box indicates the median for that group.

References

Fortier, P. A. (1989). 'Some Statistics of Themes in the French Novel'. *Computers and the Humanities.* 23 (4/5): 293-299.

Ide, Nancy M. (1989). 'A Statistical Measure of Theme and Structure'. *Computers and the Humanities.* 23 (4/5): 277-283.

Louwerse, Max M. (2004). 'Semantic Variation in Idiolect and Sociolect: Corpus Linguistic Evidence from Literary Texts'. *Computers and the Humanities.* 38 (2): 207-221.

Meyer, R. N. (1910). 'Bedeutungssysteme'. *KZ.* 43 (4): 359.

Williams, R. (1958). *Culture and Society: 1780 – 1950.* New York: Columbia University Press.

Comparing the Similarities and Differences between Two Translations

Lucic, Ana
alucic2@illinois.edu
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Blake, Catherine
clblake@illinois.edu
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

1. Abstract

Burton Pike described Rainer Marie Rilke's style in Rilke's only prose novel, *The Notebooks of Malte Laurids Brigge*, as "… arresting, haunting, and beautiful, but it is not smooth. His style is explicit, direct, almost laconic, and it has an edge." (2008, p. xvii) Pike argues that this "edge" was not sufficiently emphasized in previous English language translations and he thus wrote a new translation. The goal of this research is to explore the degree to which automated text analysis tools can capture the different styles used by Burton Pike and Stephen Mitchell in their respective translations. We are particularly interested in what kinds of similarities and differences can be captured between two renderings of Rilke's novel and in the implications of these findings on the reviews, reader reception, and critical analysis of the original work in two translations.

2. Method

Two candidate analysis tools were used to identify similarities and differences between the Pike and Mitchell translations of Rilke's novel *The Notebooks of Malte Laurids Brigge*. The first approach used a syntactic representation of the texts which was generated using the Stanford Lexical parser (`http://nlp.stanford.edu/index.shtml`). Before running the parser, each translation was tokenized into sentences using the Natural Language Toolkit tokenizer (`http://www.nltk.org/`). The generated lexical dependencies, which capture grammatical relations between words in a sentence, were then uploaded to the oracle database for subsequent analysis.

The second approach used a statistical approach–principal component analysis–to compare the manifestations of the novel. Our experiment reflects McKenna et al.'s study (1999) of similarities and differences between Samuel Beckett's French and English translation of Molloy. Three matrixes were produced comprising the 99 most common words in each of Pike and Mitchell's translation and the original text. The rows of the matrix reflect 8 text blocks of 7,500 words. Principal component analysis, which measures the variance of 99 most frequent words in the 8 blocks, was then applied to each matrix and the top eigenvectors produced were mapped into a two dimensional space for visual analysis.

3. Results

Dependency grammar analysis

Table 1 provides the summary statistics of grammatical relations which showed the largest difference captured in the parser for each translation. Although the two translations show a remarkable degree of similarity in the frequency and type of grammatical relations they employ, the dependencies revealed several differences between the two renderings. The main areas of difference were found in the use of negation modifiers, prepositional modifiers, object of preposition, parataxis, and in the word choices for adjectival and adverbial modifiers.

The dependency grammar results show that Pike used many more negation modifiers than Mitchell, which is confirmed by word frequency analysis. Pike uses the non-contracted verb, which places more emphasis on negation and thus also stays closer to the original text whereas Mitchell rarely uses a non-contracted verb form. The word "not" is not the only word in Pike's translation which indicates a negation that is used with higher frequency. Pike uses "no" with higher frequency than Mitchell, and also "nothing," "never," and "none."

The frequencies of prepositional modifier and object of preposition relations in two texts indicate a difference in how the prepositions are used in the text. This result shows a higher overall frequency of prepositions in Mitchell's translation (Table 1). The sentence level analysis of two texts, however, revealed that Pike is more likely to repeat and thus emphasize the same preposition throughout the sentence whereas Mitchell is more likely to leave out the preposition rather than repeat it. This finding suggests that in addition to the overall difference in the frequency of prepositions captured through the parser, there are indicators that the use of prepositions, their placement and distribution throughout the sentence, show differences in two texts.

The comparison of two translations revealed a higher frequency of sentences that use a semicolon to separate sub-clauses in Mitchell's translation (parataxis). Mitchell frequently arranges independent sentences using semicolons and colons, and in this way follows closely the original text, while Pike occasionally intersperses "but" and "and" in place of semicolons and colons. This finding is supported by the frequency of conjunction "but" in two translations which is found at higher frequency in Pike's translation, 559, than in Mitchell's, 506. This difference in sentence structure may provide the "edge" which Pike claims was missing in previous translations.

An examination of the unique adverbs and adjectives used in each translation revealed that although Mitchell and Pike may use similar grammatical relations in their respective renderings, their word choices are frequently different. This suggests that a semantic rather than the syntactic analysis may reveal additional differences.

PCA analysis

Figure 1 shows the score plots of the translations and suggests that the distribution of frequent terms is very similar between the two translations. However, the grouping of the 7th block with 4th, 5th, and 6th rather than with 8th block, in Mitchell's translation, calls for a closer analysis of the 7th block of the novel and its comparison with the 7th block in Rilke's original text and Pike's translation. The last part of the novel, which corresponds to the 8th block of text, is visibly different in style and tone from the rest of the novel and this difference is indicated by the location of the 8th block in the far right corner of the plot in each version.

4. Conclusion

Our results thus far suggest that syntactic grammatical relations reveal differences between the two translations that are not captured when using a bag-of-word approach (word frequencies of function words and 1,000 most frequent words). In contrast, the PCA analysis using matrices of frequent terms in the 8 blocks of text suggests that only small differences exist between the translations, with the exception of the 7th block of text, which warrants further investigation.

The similarity captured between translations using the PCA analysis brings to mind Jan Rybicki's article "Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations" in which Rybicki analyzes

the distinctiveness of character idiolects in Henryk Sienkiewicz's trilogy and concludes that: "… the patterns of difference and similarity are almost mysteriously preserved in the translations—so well that the above-mentioned linguistic differences might be the sole reason for the small differences between the original and the translation. In the greater picture, characters differ one from another in the translations just as they do in the original." (Rybicki, 2006, p.102)

The differences between two translations that were established will help create the linkages between these findings and the reader reception, reviews, and critical analysis of two translations. We also hope that they will help trace the contours along which the creation of the new variant and new literary rendering begins to emerge and the differences that speak directly to the explicit, direct, and laconic style of Pike's translation.

We plan to extend this work to include M. D. Herter Norton's translation (1949) of *The Notebooks of Malte Laurids Brigge* and to investigate how well this method generalizes to different translations of different genres.

Type of Dependency	Burton Pike	Stephen Mitchell	Dependency frequency difference	Dependency frequency difference from the mean
Negation modifier	570	242	328	164.0
Prepositional modifier	6,027	6,250	223	111.5
Object of preposition	6,034	6,254	220	110.0
Parataxis	277	496	219	109.5
Determiner	5,234	5,444	210	105.0
Possession modifier	1,408	1,571	163	81.5
Open clausal complement	1,119	1,275	156	78.0
Adverbial modifier	4,942	5,081	139	69.5
Adjectival modifier	3,016	3,153	137	68.5
Noun subject	7,441	7,559	118	59.0
Direct object	2,759	2,859	100	50.0
Clausal complement	1,692	1,617	75	37.5
Phrasal verb particle	789	734	55	27.5
Prepositional complement	308	256	52	26.0
Adverbial clause	542	593	51	25.5
Copula	1,175	1,130	45	22.5

Table 1 - Dependency distributions with the largest difference between the Pike and Mitchell translations of *The Notebooks of Malte Laurids Brigge* by Rainer Maria Rilke

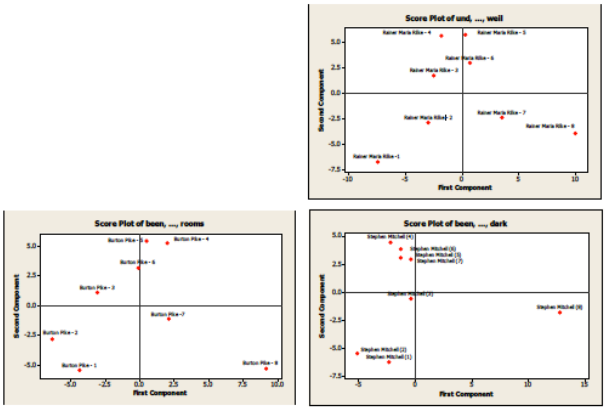


Figure 1 - Score plots

References

de Marneffe, M. & Manning, D. C. (2008). *Stanford typed dependencies manual.*. `http://nlp.stanford.edu/software/dependencies_manual.pdf` (accessed 30 October 2010).

Lucic, A. (2010). 'Measuring Similarities and Differences between Two Translations.'. *Research Showcase 2010 at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.* Champaign, IL, 2010, pp. .

McKenna, W., Burrows, J., Antonia, A. (1999). 'Beckett's Trilogy: Computational Stylistics and the Nature of Translation.'. *RISSH.* 35: 151-71.

Rilke, R.M. (2000). *Die Aufzeichnungen des Malte Laurids Brigge. [Plain Text UTF-8].* `http://www.gutenberg.org/ebooks/2188` (accessed 30 October 2010).

Rilke, R.M. (1949). *The Notebooks of Malte Laurids Brigge (Translated by M. D. H. Norton)*.. New York: W.W. Norton & Company.

Rilke, R.M. (1982). *The Notebooks of Malte Laurids Brigge (Translated by S. Mitchell)*.. New York: Random House.

Rilke, R.M. (2008). *The Notebooks of Malte Laurids Brigge (Translated by B. Pike)*.. Champaign: Dalkey Archive Press.

Rybicki, J. (2006). 'Burrowing into translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations.'. *Literary & Linguistic Computing.* 21(1): 477-495.

Digital Image Analysis and Interactive Visualization of 1000000 Manga Pages

Manovich, Lev
manovich@ucsd.edu
University of California, San Diego

Huber, William
whuber@ucsd.edu
University of California, San Diego

Douglass, Jeremy
jeremydouglass@gmail.com
Software Studies Initiative, Calit2

Digitization efforts by many museums, libraries and other institutions, and the massive growth of both user-generated and professional digital content opens us new possibilities for the study of media and visual cultures. To explore these possibilities, in 2007 we set Software Studies Initiative at University of California, San Diego. Since 2009, our research has been focused on exploring manga culture — specifically, 883 manga titles comprising 1 million digitized pages that were available as scanlations on onemanga.com in the Fall 2009. We have downloaded this whole image set along with the user-assigned tags indicating the genres and intended audiences for these titles and begun its analysis using digital image analysis, visualization, statistics, and data mining techniques.

Our paper presents our methods and some of the key findings of this ongoing research. The discussion is framed by three larger issues: 1) What are the new possibilities for describing visual language of manga made possible by using digital image analysis? 2) What do we learn by combining the software-assisted analysis of individual manga pages and the analysis of user-assigned tags? 3) What does the analysis of large cultural data sets such as 1 million manga pages tells us about cultural categories such as style and genre?

As an example of (1), we present 2D visualizations that show sets of manga pages organized by their visual characteristics measured by software (for instance, the presence of texture/detail and the amount of black in a page). Such visualizations allow us to compare visual language of individual titles, groups of titles, all titles by a particular artist, or 1 million manga pages and large sets of other kinds of images. As another example, we show visualizations that show how visual language of titles changes during the duration of their publication.

To research the relations between visual language, genre and gender/age categories in manga market as manifested in our sample (i.e. issue 2), we analyzed connections between 35 user-assigned tags (4 for audience type — shoujo, shounen, josei and seinen — and 31 for genres) available for 883 titles. We found that all tags form a connected network — i.e. any two tags occur together at least once. While some genre tags are more likely to be assigned to a particular audience segment, none of these genre tags are exclusive to male or female audiences. These empirical findings support two ideas. First, rather than thinking of "action," "adventure," "romance," or any other genre category as a distinct *genre*, we should instead understand them as genre *traits* that, according to the perception of manga fans, can be combined in a single title. Second, as constructed by these genre traits, the gender categories female / male strongly overlap.

Digital image analysis and visualization of manga pages supports the similar conclusions. If we organize 1 million manga pages by their visual characteristics and then select all pages corresponding to shoujo and shounen titles, we find that these two sets of pages form overlapping fuzzy "clouds." In other words, while large proportions of pages have distinct visual characteristics that identify them as belonging to shoujo or shounen category, a significant percentage of pages have an "androgynous" visual language.

Finally, we discuss how the analysis of 1 million manga pages leads us to to rethink the concept of style. Consider visualization which shows ourcomplete set of 1 million pages. The pages in the bottom part of the visualization are the most graphic (they have the least amount of detail). The pages in the upper right have lots of detail and texture. The pages with the highest contrast are on the right, while pages with the least contrast are on the left. In between these four extremes, we find every possible stylistic variation. This suggests that manga visual language should be understood as a continuous variable.

This, in turn, suggests that the very concept of style as it is normally used maybe problematic then we consider large cultural data sets. The concept assumes that we can partition a set of works into a small number of discrete categories. However, if we find a very large set of variations with very small differences between them (such as in this case), it is no longer possible to use this model. Instead, it is more appropriate to use

visualizations and/or mathematical models to describe the space of possible and realized variations.

Expressive Power of Markup Languages and Graph Structures

Marcoux,Yves
yves.marcoux@umontreal.ca
Université de Montréal, Canada

Sperberg-McQueen, Michael
cmsmcq@blackmesatech.com
Black Mesa Technologies

Huitfeldt,Claus
claus.huitfeldt@uib.no
University of Bergen, Norway

1. Extended Abstract

The problem of overlapping structures has long been familiar to the digital humanities community. Early on, the purely hierarchical structure has been identified as an important shortcoming of XML for the representation of meaningful features of complex works in the humanities. For example, the verse and line structures of a poem do not necessarily nest properly, but may overlap. To capture that complexity, it is necessary to host both structures in the same digital object, something which pure XML cannot do directly.

It is customary to represent the structure of a marked-up document by deriving from it a graph in which the adjacency relation among nodes corresponds to the embedding relation among the marked-up elements of the document. It is well known that, through such a construction, XML documents correspond to the subclass of graphs known as *trees*.

Numerous approaches to extend or specialize XML or SGML to handle overlapping structures in various ways have been proposed in the literature over the years (see for instance [B1995]). Some of these approaches work on the graph-structure view of XML. The graph-structure is no more required to be a tree, but it must still satisfy basic constraints (e.g., finiteness, non-circularity) that make them plausible models for “documents.” One such proposal is the GODDAG (*General Ordered-Descendant Directed Acyclic Graph* [SH2004]). Roughly, GODDAGs are like XML trees except that they allow multiple parenthood and do not require a total ordering on leaf nodes. (Thus, XML trees constitute a subset of GODDAGs.)

Other proposals address the problem by working on the markup language side. In this group, we find pure XML solutions (which represent the non-hierarchical structures through coding conventions and a corresponding ad hoc layer of semantics, e.g., milestones, fragmentation, virtual elements, etc. [B1995] [SH1999] [W2005]) and non-XML ones, based on *generalizations* of XML allowing non-embedding constructs, such as overlapping elements. Among the latter is TexMecs [HS2003], an XML-like markup language allowing overlapping elements (as well as other constructs).

An interesting question is how the various proposals compare with respect to their expressive power. Beyond a few obvious subset/superset relations that exist among the formalisms, very little is known in this area. In 2008, Marcoux established a result [M2008] linking a subclass of GOODAGs to a subset of TexMecs. *Overlap-only TexMecs* (or *oo- TexMecs*) is the subset of TexMecs that uses only embedding and overlapping elements (and none of the other constructs). A graph whose structure corresponds to an oo-TexMecs document is said to be *oo-serializable*. Marcoux showed (essentially) that the oo-serializable GODDAGs are exactly the *completion-acyclic* ones, thus showing that oo-TexMecs and the class of completion-acyclic GODDAGs have the same expressive power.

In this paper, we extend that result in two ways. First, we define *child-ordered directed* graphs (CODGs), a class of graphs strictly larger than that of GODDAGs, and argue that, in spite of its generality, it still captures a plausible and interesting notion of "document." Then, we show that the strictly stronger expressive power of the CODG, compared to the GODDAG, vanishes when oo-serializability is required. This constitutes a strong indication that overlap-only markup languages may be insufficient for representing complex document structures.

More precisely, we show that the classes of single-root CODGs and GODDAGs coincide when restricted to completion-acyclic graphs. There are, however, completion-acyclic multiple-root CODGs that are not oo- serializable. We show that, for multiple-root CODGs, the stronger condition of *full-completion-acyclicity* characterizes oo-serializability.

The definition of fully-completion-acyclic graph does not in itself suggest an efficient algorithm for checking the condition, nor for computing a corresponding overlap-only document when the condition is satisfied. We present ideas that could be exploited to accomplish those tasks efficiently.

The main conclusion of this work is that markup languages that generalize XML only by allowing overlapping elements may not be expressive enough to represent complex document structures. Indeed, our results show that the requirement that a graph be serializable in an overlap-only language effectively prevents the presence of too complex structures in the graph. The new graph structure introduced in the paper, the CODG, is of interest in itself as an abstract document model. Finally, the ideas we present on how to detect whether a graph is serializable using only overlapping elements and on how to then compute a serialization would be helpful in authoring environments, because they would allow serializing documents using the most simple constructs (e.g., overlapping elements) whenever possible, while resorting to more complex constructs (e.g., virtual elements) only when absolutely necessary.

References

Barnard, David, Burnard,Lou , Gaspart, Jean-Pierre, Price, Lynne A., Sperberg-McQueen, C. M., Varile, Giovanni Battista (1995). '"Hierarchical Encoding of Text: Technical Problems and SGML Solutions"'. *Computers and the Humanities.* 3: 211-231. http://www.springerlink.com/content/p7775247276v88h3/http://xml.coverpages.org/barnardHier-ps.gz.

Huitfeldt, Claus, Sperberg-McQueen, C. M. (January 2001, rev. October 2003). *TexMECS: An experimental markup meta-language for complex documents.* University of Bergen. http://mep.blackmesatech.com/mlcd/2003/Papers/texmecs.html.

Marcoux, Yves (12-15 august 2008). 'Graph characterization of overlap-only TexMECS and other overlapping markup formalisms'. *Proceedings of the Balisage 2008 conference.* Montréal (Canada). http://www.balisage.net/Proceedings/vol1/html/Marcoux01/BalisageVol1-Marcoux01.html.

Sperberg-McQueen, C. M., Huitfeldt, Claus. *GODDAG: A Data Structure for Overlapping Hierarchies*. Springer-Verlag (2004).

C. M. Sperberg-McQueen, Claus Huitfeldt. '"Concurrent Document Hierarchies in MECS and SGML"'. *Literary and Linguistic Computing* , 14 1999, pp. 29-42. http://llc.oxfordjournals.org/cgi/content/abstract/14/1/29.

Witt, Andreas (2005). *"Multiple Hierarchies: New Aspects of an Old Solution". Heterogeneity in*

Focus: Creating and and Using Linguistic Databases. Interdisciplinary Studies on Information Structure (ISIS), Working Papers of the SFB 632.. Stefanie Dipper, Michael Götze, Manfred Stede (eds.). Germany: University of Potsdam V. vol. 2. http://www.sfb632.uni- potsdam.de/publications/isis02_4witt.pdf.

Omeka in the Classroom: The Challenges of Teaching Material Culture in a Digital World

Marsh, Allison
allisonmarsh@yahoo.com
University of South Carolina

For presenters and attendees at DH2011, there is no need to sing the praises of the digital world. We are the early adopters, the converted, the evangelists. But our colleagues across the humanities are not yet entirely convinced, and of more concern to me, neither are the students. I direct the museum studies track of the masters in public history at the University of South Carolina, one of the oldest public history programs in the country. It is a nationally competitive program, and our graduates have an impressive placement record: the Smithsonian; the National Park Service; federal, state, and local government. And yet, since I joined the faculty three years ago, I have been shocked that the students – the so-called digital natives – have little interest in the digital world as part of their professional training. They may communicate with each other using Facebook, share photos on Flickr, or post to their personal blogs, but when it comes to coursework they expect, and sometimes demand, a traditional graduate seminar where we read and discuss books. More than one student has balked at my assignments, whining, “I don’t need to learn how to program. I just want to be a *regular* historian.” Unyielding in my persistence, I argue back that it is no longer an option. Wikis, blogs, and tweeting are everyday realities for museum professionals. At the very minimum, all curators and collections managers need to have a basic understanding of database architecture in order to structure their object databases and construct useful queries. More importantly, two decades of digitization has created new questions for curators of three-dimensional objects: What does material culture look like on the web? How do you curate it? How does the public interact with virtual objects? What is the relationship between virtual and physical museum artifacts?

Each fall I teach HIST 787: Material Culture Studies, the foundational graduate seminar for the museums track in our masters program. On the first day of class, I ask the students to bring in five objects that describe either themselves or their research interests. I tell them to choose wisely, as they will be using those objects

every week for the entire semester, but otherwise I give no guidance to object selection. The objects serve multiple purposes throughout the semester, but most importantly they are part of a larger project to create an object database that represents the changing attitudes towards material culture in the digital age. Each year the students must create an online exhibit drawing from the objects in the database, the objects of both their classmates as well as the students of previous years. Clearly each year the number of objects in the database increases, and the distance from the early contributors becomes greater. I am in Year 3 of what I anticipate to be a decade long study, and this short paper presentation is designed to give preliminary results. Because this course is part of a two-year masters degree, this is the first year where the students do not have direct access to the owners of objects from previous classes.

I have chosen Omeka as the platform for this assignment. Although I am well aware of the limitations, as well as the potential, of the open source software, Omeka has a low barrier for entry. Omeka was developed by George Mason University's Center for History and New Media with non-IT specialists in mind. CHNM's goal was (and continues to be) to provide museum and library professionals with a tool that allows them to concentrate on content and interpretation without worrying about programming. I am concerned that by using a black box application my students don't fully understand the implications of engaging with the virtual world. However, that is one of the compromises I have made in order to encourage budding historians to get their toes wet in the digital arena.

For their final assignment, students must create records for each of their objects, which includes uploading images, entering Dublin Core metadata, tagging objects with key words, and writing short descriptions. The students then must curate their own exhibit, either by using one of the theme templates provided by CHNM or by creating their own. The open source software allows students who are more skilled or interested in web design to create more elaborate exhibits.

So far the results of the online exhibits have been mostly disastrous. As a whole, the exhibits are terrible (available at `http://hist787.cas.sc.edu`). They have clunky navigation, lack any elegance in design, and often are just plain boring. In many ways, the exhibits are proof of my distrust of black box software for developing online exhibits and are an indicator that anyone who wants to engage seriously in the virtual world needs significantly more training (either formally or informally) than a few hours of online tutorials can provide. More generously, these online exhibits are often the first experience students have in curating, and so one of the assignment's goals is for students to gain skills in developing effective narrative techniques (useful in both physical and digital curation). In assessing their work, it is important to be mindful of the learning process; remember that they are professionals in training, and they should not be judged on their first attempt but rather on the progress they achieve by the time they graduate.

However, as a pedagogical device, the assignment has been tremendously successful. By working through the process of creating an online exhibit, the students naturally confront the many epistemic questions relating to the use of physical objects in a virtual environment. Students immediately recognize the diverse challenges of working in the digital format, from the pedestrian, such as how to search for items when a previous user failed to enter appropriate metadata, to the substantial, such as questioning the ethics of using an object as a metonym in an exhibit that is antithetical to the physical object's authenticity. My goal for the assignment is not for students to become master web designers, but for them to engage in the questions confronting digital curation.

Although I could talk at length about the implications of this ongoing assignment, this short paper will focus on the joint challenges of curating digital resources and the role of digital humanities in the academic curricula – how are universities training the next generation of museum professionals who will have to confront digital curation. My presentation will be a snap shot of a longitudinal study that is currently in progress. I will briefly describe the assignment, its goals, and the unexpected lessons learned thus far. I hope to reach fellow members of the academy to discuss effective teaching techniques while at the same time seeking feedback from museum professionals as to what skills they believe graduating students should have. How do professors balance the need to provide theoretical training in how to read and interpret material culture while fostering the development of technical skills in an ever-changing digital landscape?

Towards a Narrative GIS

McIntosh, John
jmcintosh@ou.edu
University of Oklahoma

De Lozier, Grant
ghaxed@gmail.com
University of Oklahoma

Cantrell, Jacob
jcantrell@ou.edu
University of Oklahoma

Yuan, May
myuan@ou.edu
University of Oklahoma

1. Introduction

Research in narrative intelligence applies artificial intelligence approaches to study human ability to organize experience into narrative form (Mateas and Sengers 2003). Narratives are traditionally defined as “a series of temporally ordered clauses” (Labov 1972, p360-361). The time-centric approach leads to a lesser consideration of space in narrative construction and analysis. In contrast, we advocate a geospatial narrative in order to stress the importance of space and time in understanding the ordering and spatial interaction of events.

We define a geospatial narrative as a sequence of events in which space and time are equally important. Narratives are stories that constitutes sequential organizations of events (Franzosi 2010). Each event in a narrative relates sequential or consequential occurrence in space and time. The conventional Geographic Information Systems (GIS) center on information about spatial states of reality, and temporal information is handled as add-ons to spatial objects. Alternatively, we conceptualize a narrative GIS that emphasizes representing and ordering events in space and time as well as functional abilities to construct meaningful geospatial narratives. While an event is a complex, fuzzy term, we start with one basic linguistic element of narratives: action, as the primitive data construct to start building a narrative GIS. By relating action events across space and time, a narrative GIS aims to discover spatiotemporal correlates among actions and relate actions across scales.

Depending on the perspectives, there are many kinds of events, e.g. instantaneous events, discrete events, cyclic events, transitional events, and others. In contrast to TimeMaps (Farrimond et al. 2008), our vision of a Narrative GIS goes beyond spatiotemporal visualization to spatial analytics. By using action events as the primitive data constructs, a narrative GIS can support spatial queries of sequential and consequential actions. A Narrative GIS is therefore capable of revealing how time unfolds change and space unfolds interactions (Massey 2005).

We use two distinctive corpuses of histories in building narrative GIS databases and narrative analytics as a proof of concepts: Dyer's *Compendium of the War of the Rebellion* and the *Richmond Daily Dispatch*. Frederick H. Dyers, a Civil War veteran compiled the *Compendium* based on materials from the Official Records of the Union and Confederate Armies and other sources. The compendium lists organizations and movements of regiment cavalries mustered by State and Federal Governments for services in the Union Armies. Collaborating with the digital scholarship group at the University of Richmond, we have started with four files from Dyer's Compendium: the 45th Massachusetts Infantry, the 107th Pennsylvania Infantry,the1stCaliforniaInfantry,andthe1stNewYorkCavalry. The *Richmond Daily Dispatch* was one of the primary news media in the south during the Civil War. The newspaper was one of the most widely distributed newspapers of the south and included news from the entire east coast. The *Richmond Daily Dispatch* retained the reputation as politically unbiased was published throughout the Civil War.

2. Methodology

Our idea of a narrative GIS consists of (1) semantic elements (who did what), (2) temporal elements (when), and (3) spatial elements (where). A geospatial narrative object integrates the three elements and enables search for and analysis of spatial and temporal relationships among narrative objects. Input data for narrative GIS vary widely from structured to unstructured sources. In this study, both input data are texts, albeit in very different structures. Dyer's Compendium concisely lists regiment movements. *Richmond Daily Dispatch* consists of news articles. Spatial and temporal connections among units in these texts are considerably different. Nevertheless, the conceptual framework of a narrative GIS demands the identifications of semantic, temporal, and spatial elements from the texts to form narrative objects and relationships. As such, our workflow includes six

key steps: (1) extract text analysis units; (2) identify action verbs; (3) identify time for words and text units; (4) identify locations for words and text units; (5) combine all identified elements into a GIS database; and (6) build spatial and temporal relationships among narrative objects. A schematic view of the workflow is presented in figure 1

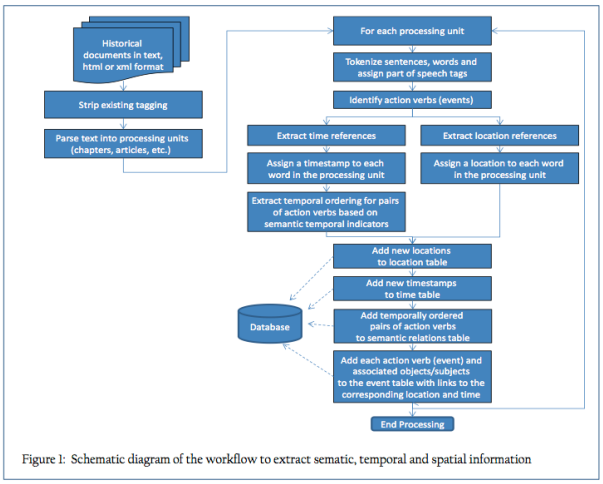


Figure 1

We begin with electronic versions of the historical documents. The texts are split into subsets such as newspaper articles or book chapters for processing. These subsets are typically written as a unit and need to be analyzed that way for successful interpretation. For each processing unit, we apply natural language processing to tokenize sentences and identify parts of speech (e.g. verbs and nouns). The parts of speech provide important clues to extract information.

The work presented here is centered on the location, time, and other characteristics of events. The part of speech tagging is used to identify verbs. We are most interested in “action” verbs and refine our list of potential candidates by removing stative and modal verbs. Location referencing begins with recognition of a standard grammatical structure to the way locations appear in text. In general, locations are proper nouns that do not directly follow a determiner (except for physical features).

Candidate words are matched to all their possible real-world geographic referents in the “Gazetteer Matching” process. A number of different gazetteers are utilized in the matching including the US Populated Places gazetteer and State hydrography datasets from the USGS, historical counties, states, and territories files from National Historical Geographic Information System, and the US Census Bureau’s historical 100 largest cities dataset (US BGN; US NHD; NHGIS 2008; Gibson 2008). These data are assembled in GIS and each location is identified with a historically and spatially appropriate hierarchy. The names of geographic locations are often highly ambiguous. For example “Georgetown” has over 70 possible locations among U.S. cities. Disambiguating a word to its true location is an important and difficult task. A substantial amount of work has already been done on location disambiguation under the heading of “Toponym Resolution” (Leidner2007;Leidneretal2003). Figure 2 illustrates the steps in the location referencing process.

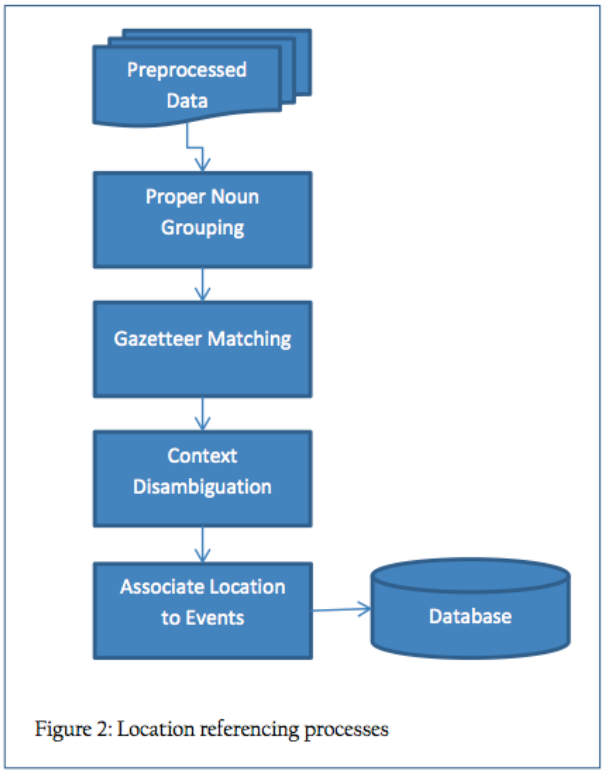


Figure 2

The temporal processing steps aim to extract dates, durations of events and the relative temporal ordering between events. Historical texts contain temporal information in a variety of formats. Most obvious are explicit dates that include information such as the day, month and year. In addition, these texts often include clues to derive dates and relative ordering of events. For example words such as “yesterday“ or “last week” allow the date to be derived based on a temporal relation to an anchor date (Han 2006). Similarly, relative temporal expressions allow explicit dates to be determined based on temporal relations to the current temporal focus (Han 2009).

Figure 3 outlines the steps in our approach.

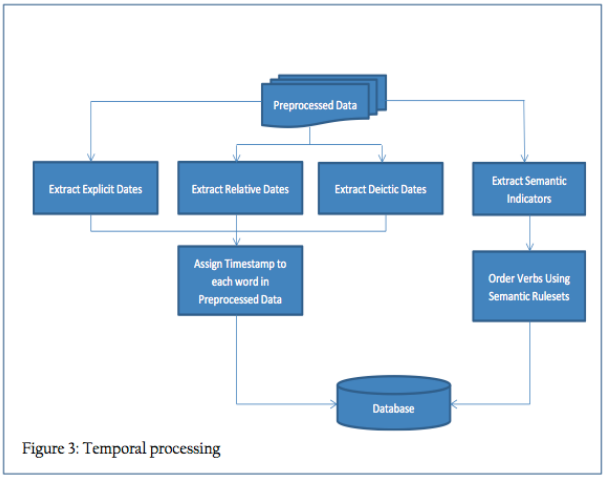


Figure 3

We begin by extracting anchor dates such as the date of publication for a newspaper article and explicit dates found in the text. We use temporal indicator words to refine the date of events and help establish temporal ordering. Explicit dates contained in the text are modified by deictic or temporal expressions. Semantic relationships between events are extracted based on semantic indicators. When all of the temporal information is relative and there are no explicit dates to give an explicit order, thirteen temporal relationships are used to find the temporal ordering (Allen 1983).

3. Results and Concluding Remarks

Thus far, our effort has been focused on extraction of events from the natural language text sources, anchoring the events to geographical locations and in time, and extracting information on the actors and objects involved in the events. Figure 4 illustrates results from Dyer's Compendium of the War of the Rebellion.

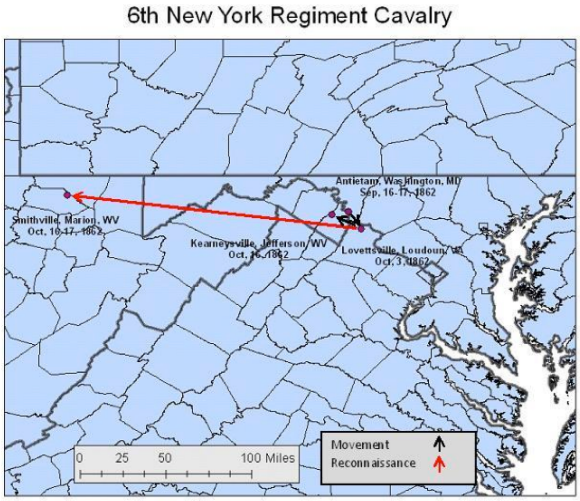


Figure 4

This figure shows the activity of the 6th New York Regiment Calvary in Maryland and Virginia in September and October of 1862. The processing identified the events including the regiment's movements and splitting off of a reconnaissance mission from Lovettsville to Smithville while the main regiment moved to Kearnysville. While this example from a single source, it illustrates the potential for the system to support more complex geospatial narratives with the addition of information from other sources.

Figure 5 shows a visualization of a Richmond Daily Dispatch article.

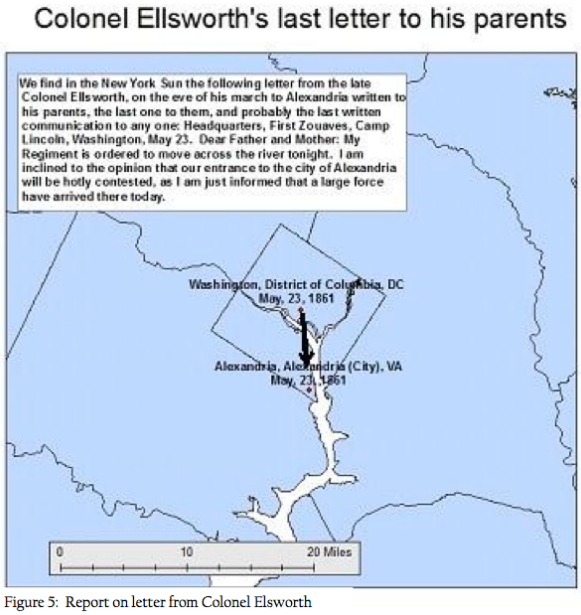


Figure 5

The article describes a letter from a Union colonel to his family. It discusses the Union's plan to move troops

to Alexandria Virginia the next evening. The article illustrates that in addition to working with events that had already occurred the approach can also be used to help investigate the thoughts and motivation leading to events that had yet to occur.

These examples of preliminary results demonstrate the basic use of a Narrative GIS. As we continue building the event narrative database, additional functions will be built in for narrative analytics. For example, we are interested in deciphering the local, regional and national processes on emancipation and to identify scalar effects on military, political, and individual processes. One approach will be extracting reports on battles and run-away slaves and analyze spatial and temporal correlations among these events. When we extract events of different categories in space and time, a Narrative GIS will allow us to analyze spatial and temporal relationships among these kinds of events to draw insights into the integration of multiple perspectives and interpretations of geospatial narratives.

References

Allen, J., Waltz, D. (1983). *Maintaining Knowledge about Temporal Intervals. Communications of the ACM.* .

Bird, S., Klein, E. , Loper, E. (2009). *Natural Language Processing with Python --- Analyzing Text with the Natural Language Toolkit.* Cambridge, MA: O'Reilly Media.

Farrimond, B., Presland, S., Bonar-Law, J. , Pogson, F. (2008). *Making History Happen: Spatiotemporal Data Visualization for Historians. Second UKSIM European Symposium on Computer Modeling and Simulation.* Liverpool, UK: IEEE.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database.* Cambridge, MA: The MIT Press.

Franzosi, R. (2010). *Quantitative Narrative Analysis.* Los Angelas, CA: SAGE Publications, Inc..

Gibson, C. (2008). *Population Of The 100 Largest Cities And Other Urban Places In The United States: 1790 to 1990. U.S. Census Bureau, Population Division.* . http://www.census.gov/population/www/documentation/twps0027/twps0027.html.

Han, B., Gates, D., Levin, L. (2006). *From Language to Time: A Temporal Expression Anchorer. Proc. 13th International Symposium on Temporal Representation and Reasoning*

Han, B. (2009). *Reasoning about Temporal Scenario in Natural Language. In the Proceedings of AAAI Workshop on Spatial and Temporal Reasoning*

Jackendoff, R. (1992). *Languages of the Mind.* Cambridge, MA: The MIT Press.

Labov, W. (1972). *Language in the inner city.* Philadelphia: University of Pennsylvania Press.

Leidner, J, L., Sinclair, G., Webber, B (2003). *Grounding spatial named entities for information extraction and question answering. Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References.* Edmonton, CAN.

Leidner, J (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Diss..* University of Edinburgh, School of Informatics. Institute for Communicating and Collaborative Systems.

Massey, D. (2005). *For Space.* Thousand Oaks, CA: Sage.

Mateas, M., Sengers, P. (eds.) (2003). *Narrative Intelligence*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Miller, G., Beckwith, R. , Fellbaum, C., Gross, D., Miller, K. (1993). *Introduction to WordNet: An Online Lexical Database (Revised).* Princeton University.

Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T. (2006). *Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation. In Proceedings of The Fifth International Conference on Language Resources and Evaluation(LREC).* http://www.arxiv.com/ftp/cs/papers/0609/0609065.pdf.

(2008). *National Historical Geographic Information System.* Minneapolis, MN: Minnesota Population Center: University of Minnesota.. http://www.nhgis.org.

U.S. Board on Geographic Names: Domestic and Antarctic Names – State and Topical Gazetteer Download Files. United States Geological Survey. http://geonames.usgs.gov/domestic/download_data.htm.

U.S. Geological Survey: National Hydrography Dataset. United States Geological Survey. http://nhd.usgs.gov/data.html.

Charlotte's Web: Encoding the Literary History of the Sentimental Novel

Melson, John
John_Melson@brown.edu
Women Writers Project, Brown University

Funchion, John
jfunchion@mail.as.miami.edu
University of Miami

Ever since the cultural turn in literary studies, literary scholarship has focused on examining the cultural work performed by texts. Indeed, one of the enduring theoretical innovations of the 1980s was the reconceptualization of literary value according to the concept of cultural work—the idea that texts should be evaluated not by the innate aesthetic or formal qualities they manifested, but according to their "designs upon their audience" and their ability "to make people think and act in a particular way" (Tompkins 1985). As a result, previously overlooked or marginalized texts have often been deemed suitable for study in the ensuing years based largely the on degree to which they are said to perform particular kinds of cultural work. In one familiar example from the arena of nineteenth-century American literature, scholars have come to see a stereotypically sentimental novel like *Uncle Tom's Cabin* as a canonical work primarily because its complex entanglements with the politics of the antebellum United States allow it to be read— to cite Tompkins once more—as part of a "monumental effort to reorganize culture." But while this approach has become an important characteristic of much literary scholarship, the question of *how* to measure a given text's capacity to perform cultural work remains vague and often myopically focused on the synchronic significance of the work in question at a specific point in time.

In this paper we investigate the matter of cultural work from the vantage point offered by current scholarship in the digital humanities. We do so with an eye toward developing a model of literary history that draws on uniquely digital methods for structuring and formalizing intertextual relationships, and that make it possible to chart the cultural and formal significance of literary works across space and time. Specifically, we use Susanna Rowson's late eighteenth-century sentimental novel *Charlotte Temple* as a case study in how digital literary history can evaluate the meaning and formal significance of a text diachronically. Often cited as one of the earliest American bestsellers, Rowson's work was reprinted hundreds of times during the nineteenth century and was referenced repeatedly in an extensive body of writing, ranging from reviews and advertisements to melodramatic stage adaptations and ostensibly factual regional histories of the United States. Tracking the novel's extended nineteenth-century afterlife through this complex network of external references, we demonstrate how the application of detailed interpretive markup to the multiple documents that reference *Charlotte Temple* (instead of the novel that would, more conventionally, be thought of as the "primary" text) enacts a theory of literary history in which concepts of cultural work may best be observed as phenomena that develop over time.

While the idea of interpretive markup is not new, it is most often used either as a means of recording a layer of scholarly annotation on some particular document or classifying portions of a document according to some external taxonomy. Both uses are supported, for instance, by the current version of the Text Encoding Initiative (TEI) Guidelines, which provides examples of specific methods for encoding "semantic or syntactic interpretation" using a set of standard TEI XML elements and attributes. In practice, though, discussions of interpretive markup in digital humanities projects often revolve around questions of readability and the appropriateness of stand-off versus inline markup (e.g. does excessive interpretive markup pose problems for human readability of encoded text? [Campbell 2002]), or issues of preservation, curation, and reliability (e.g. what challenges for encoding textual information accurately and reliably are created by allowing extensive interpretive markup? [Berrie et al. 2006]). Although such questions are important in certain contexts, our project treats them as less important than the question of how interpretive markup may be enlisted as a surface for scholarly analysis—that is, how interpretive structures can themselves be classified and interpreted.

Our project adapts several of the interpretive structures provided in TEI XML to represent the connection between specific formal properties in *Charlotte Temple* and external evidence of the novel's broader cultural influence, as attested by external references to it. We record basic metadata about each external reference (author, title, date and location of publication, etc.) while also indicating which specific features of the novel the reference comments on, as well as the nature of that commentary. The result is a set of XML-encoded

documents classified according to the interpretive work they do: a record, as it were, of how nineteenth-century readers interpreted and responded to the novel's formal properties. At the same time, a secondary layer of interpretive markup further formalizes the relationships we identify across this primary layer of interpretation. Taken together, both layers provide substantial material for further abstraction: for instance, the automated generation of topic maps that represent a variety of cultural knowledge structures. The result is a web of connections in the form of citations, references, allusions, parodies, and comments spanning the nineteenth century, whose formalized structures constitute the interpretive tissue of digital literary history. The ability to map and visualize these structures, we argue, offers significantly new possibilities for observing cultural work as a phenomenon that evolves over time, and whose connection to particular texts is most productively understood as operating at the nexus of close reading and "distant reading" strategies.

Our work constitutes an initial small-scale experiment, but it has relevance for larger questions of scale, purpose, and method in both the digital humanities and conventional modes of literary scholarship. In particular, it suggests that although digital scholarship offers a potential reconceptualization of literary history, practices of literary history also pose interesting challenges for work in the digital humanities. In recent years, projects in the digital humanities have increasingly responded to the question of scale by refining methods for aggregating, classifying, and analyzing enormous bodies of textual material—in other words, by treating text as a mass of data that can yield meaningful responses to statistical analysis of its language. Whether offering answers to now-familiar questions about scale itself— "What do you do with a million books?"—or taking up Franco Moretti's challenge to peer into the "cellars of culture" represented by the tens of thousands of unread and unknown novels published in the nineteenth century, text mining and other forms of "computational humanities" have been increasingly held up as a means of "checking the generalizations of literary history"(Crane 2006; Parry 2010; Pasanek and Sculley 2008). At the same time, long-running digital humanities projects with deep investments in scholarly text encoding have tended to approach literary history from the opposite direction, emphasizing how "craft encoding," for instance, employs editorial methodologies and modes of textual scholarship that participate in "making literary history" in the digital medium (Flanders 2009; Dimock 2007). While neither approach directly contradicts the other—indeed, in practice they often coexist harmoniously—they inflect the concept of literary history in crucially different ways. The former often treats the machine-processable language of texts as a primary axis along which historical change in writing manifests itself: intertextual relationships are revealed in changing patterns of linguistic borrowing, with the emphasis on facilitating the "comparability of texts" (Mueller 2008). The latter, while still valuing intertextual comparison, tends to prioritize the documentary and contextual over the linguistic—for instance recording textual variation across editions of the same text, or formalizing through encoding "rich environmental contextualizations" for the study of particular texts (Folsom 2007). Our project borrows from both strategies, and in doing so seeks to demonstrate in practice a method by which digital literary history negotiates this apparent bifurcation.

References

Berrie, Phill et al. (2006). 'Electronic Textual Editing: Authenticating Electronic Editions'. *Electronic Textual Editing.* Burnard, O'Keefe, and Unsworth (ed.). New York: MLA. http://www.tei-c.org/About/Archive_new/ETE/Preview/eggert.xml. .

Crane, Gregory (2006). 'What Do You Do with a Million Books?'. *DLib Magazine.* [. http://www.dlib.org/dlib/march06/crane/03crane.html] .

Dimock, Wai Chee (2007). 'Introduction: Genres as Fields of Knowledge'. *PMLA*. , pp. 1377-88.

Flanders, Julia (2009). *2009 TEI annual conference*, Ann Arbor, Michigan. Seminars in Scholarly Text Encoding with TEI. Paper presented at the http://www.wwp.brown.edu/encoding/research/tei2009/presentations/html/TEI 2009_lecture.xhtml.

Folsom, Ed (2007). 'Database as Genre: The Epic Transformation of Archives'. *PMLA*. , pp. 1571-9 . .

Mueller, Martin (2008). 'TEI-Analytics and the MONK Project'. *2008 TEI annual conference* Paper presented at the , London, UK. http://www.cch.kcl.ac.uk/cocoon/tei2008/programme/abstracts/abstract-169.html.

Parry, Marc (28 May 2010). 'The Humanities Go Google'. *The Chronicle of Higher Education* . [http://chronicle.com/article/The-Humanities-Go-Google/65713/] .

Pasanek, B. and D. Sculley (2008). 'Mining Millions of Metaphors'. *Literary and Linguistic Computing*Literary and Linguistic Computing , 23 (3): 345-60. .

Text Encoding Initiative. P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AI.html.

Tompkins, Jane (1985). *Sensational Designs: The Cultural Work of American Fiction, 1790-1860.* New York: Oxford University Press.

Wilkens, Matthew (2009). 'Corpus Analysis and Literary History'. *Digital Humanities 2009*Paper presented at , College Park, MD. .

The Digital Dictionary of Buddhism: A Collaborative XML-Based Reference Work that has become a Field Standard: Technology and Sustainable Management Strategies

Muller, Charles. A.
acmuller@jj.em-net.ne.jp
Center for Evolving Humanities, University of Tokyo

The Digital Dictionary of Buddhism (DDB) (http://buddhism-dict.net/ddb/subscribing_libraries.html), now on the Web for more than 15 years, has come to be regarded as a primary reference work for the field of Buddhist Studies. Containing over 54,000 entries, it is subscribed to by more than 38 university libraries (http://www.buddhism-dict.net/ddb/subscribing_libraries.html). It is supported by the contributions of over 70 specialists, many of these recognized leaders in the field. It can perhaps be described as example of the type of web resource hoped for by Jaron Lanier in his book *You Are Not A Gadget*, or something similar to the sort of thing envisioned by Joseph Raben in his Busa lecture at the DH2010, in the sense of being the fruit of the collaborative efforts of a community of scholars that has reached a degree of status and sustainability such that it has been able to grow and thrive—despite having little funding or the support of a major organization or team of programmers—in the age where such resources are so readily eclipsed by the combination of Wikipedia and Google. The field of Buddhist Studies has its own reliable, scholarly-edited, fully documented and responsible online reference work that has developed a center of gravity sufficient for it to continue to grow as the resource that specialists turn to without hesitation, and to which they may contribute knowing that they will be clearly accredited, and that what they write will not be deleted or changed in the following moment by, for example, a junior high school student.

With the DDB having a history almost equal in length to that of the WWWeb as we know it (it went online in 1995), there is a wide range of issues that can be discussed beyond its present technical structure. Of great importance are the management strategies that have allowed its continued progress through the

long series of changes the Web has witnessed during this first epoch of the Internet. How, exactly, can a project that is based on the continual development of quantity and quality of reference data can continue to grow to the extent of becoming the de facto primary field reference work after exhausting the first couple initial grants, without either becoming a fully “pay-for” resource (perhaps being bought out by a commercial enterprise of some sort), or being supported by some private organization — an alternative fraught with the danger of forcing the resource to co-opt its principles and its objectivity?

I began the compilation of the DDB and its companion CJKV-English Dictionary (CJKV-E) in 1986, originally simply envisioning the eventual publication of the usual printed work. In 1994, however, the Web made its appearance, and the potential advantages of trying to develop a reference work online in a collaborative manner were immediately apparent. So in the middle of 1995, I converted my WordPerfect word-processor files to HTML, and placed the dictionary on the web. To my great elation, I was soon contacted by a few good scholars with similar interests, who were willing to offer both content and technical advice.

During its first few years on the web, the DDB was maintained in a simple, hard-linked HTML format. With the help of Christian Wittern, this source was converted to SGML, and then XML. A major turning point in the history of the project came in January 2001, when Michael Beddow offered to help with Web implementation, and for the first time, the raw XML data was searched and presented to users through a combination of Perl and XML/XSLT technology. At that time, building a search engine that could deal with mixed Western/CJK text in UTF-8 encoding was a not at all a simple matter, so Michael's search engine was a bit of a novel creation—and was able to serve its purpose with only minor tweaks up through most of 2010, for almost a full decade.

When Michael Beddow's search engine was set up in 2001, usage of the DDB increased dramatically. Yet despite our repeated pleas for user contributions, except for a very small number of “enlightened” individuals who somehow naturally grasped the meaning of this strange new thing called “web collaboration,” it became apparent that there were very, very few people willing, on their own, to take five or ten minutes to write up and send us even a couple of terms from their own research work. This lack of interest on the part of users in making contributions was extremely disappointing. Thus, while our password security system was originally set up to ward off hacking attempts, we decided to experiment with using this apparatus to institute a two-tiered system of access. In the first level, any user could access the data a limited number of times in a 24-hour period, logging in as *guest*. In the second level, contributors were granted unlimited access. We started off setting the *guest* limit at fifty, but leaving it at this amount for a few weeks we received neither complaints nor contributions. We then began to gradually drop the number down to forty, thirty, and then twenty searches in a day. At twenty, there was still nary a complaint made nor contribution to be seen. But when we hit the number of ten, everything changed. We were first bombarded with indignant complaints, but holding the line, and at the same time lowering the minimum required level of contribution to the equivalent of one A4 page for two years of access, eventually these complaints began to turn into contributions. This was a watershed moment for the project, because we found that once people contributed one time, most of them continued to do so, whether voluntarily, or by continued prompting through this same arrangement.

At the time of my first public presentation of the DDB at a meeting of the Electronic Buddhist Text Initiative (EBTI http://buddhism-dict.net/ebti/) in 1996, the DDB contained approximately 3,200 entries. That number is now over 54,000, with a present average growth rate of 4,000 terms per year. The continued growth in popularity of the DDB, especially as a reference work for graduate and undergraduate courses in Buddhist Studies in North America and Europe generated one more access problem that needed resolution—that of how to allow for the use of the DDB in the case where an instructor wanted to use the dictionary for an university course. To deal with these kinds of situations, we decided to begin to offer subscriptions to university library networks for a modest fee. This policy brought about an unforeseen benefit, in that we could now provide a list of reputable institutions that had deemed the DDB to be an academic reference of high standards. It also generated a small but steady income, which allowed us to pay for hardware and software, and a couple of part-time workers to do input and editing. Finally, in order to encourage contribution from qualified scholars, great effort was expended toward letting members of the field know of the contributions being made by their colleagues. Thus on the dictionary's web site itself, as well as on associated news and mail lists, information regarding new contributions is energetically distributed.

This presentation will start off with a short demonstration of the most advanced functions of the DDB, to be followed by a brief overview of its

technical framework (P5- influenced XML, delivered through XSL and Perl). We will then outline the above-introduced key factors of the management of the DDB that we believe have most directly contributed to its great success.

References

Lanier, Jaron (2010). *You Are Not A Gadget.* New York: Alfred A. Knopf.

Muller, A. Charles (2009). 'The Digital Dictionary of Buddhism [DDB]: Present Status and Future Developments.'. *Scholars of Buddhism in Japan: Buddhist Studies in the 21st Century.* Kyoto: International Research Center for Japanese Studies, pp. 87–100.

Muller, A. Charles, and Michael Beddow (2002). 'Moving into XML Functionality: The Combined Digital Dictionaries of Buddhism and East Asian Literary Terms'. *Journal of Digital Information: Special Issue on Chinese Collections in the Digital Library.* Volume 3, issue 2. http://journals.tdl.org/jodi/article/view/jodi-65/82.

Raben, Joseph (2010). 'Humanities Computing in an Age of Social Change'. *DH2010.* July 8, 2010http://www.artshumanities.net/video/roberto_busa_award_lecture_joseph_raben_–_humanities_computing_age_social_change_dh2010.

Tasks vs. Roles: A Center Perspective on Data Curation Needs in the Humanities

Muñoz, Trevor
munoz14@illinois.edu
Center for Informatics Research in Science and Scholarship, University of Illinois, Urbana-Champaign, USA

Varvel, Virgil
vvarvel@illinois.edu
Center for Informatics Research in Science and Scholarship, University of Illinois, Urbana-Champaign, USA

Renear, Allen H.
renear@illinois.edu
Center for Informatics Research in Science and Scholarship, University of Illinois, Urbana-Champaign, USA

Trainor, Kevin
trainor1@illinois.edu
Center for Informatics Research in Science and Scholarship, University of Illinois, Urbana-Champaign, USA

Dolan, Molly
molly.dolan@mail.wvu.edu
West Virginia University, USA

1. Abstract

To support the development of curricular content for the Data Curation Education Program (DCEP) at the Graduate School of Library and Information Science (GSLIS), University of Illinois at Urbana Champaign, a needs analysis case study focusing on digital humanities centers was carried out in late 2010. Collectively the results paint an interesting picture of the perception of humanities data curation needs by directors and key staff. Several results were contrary to what we anticipated; for instance, there was only modest agreement on critical areas of expertise needed to sustain meaningful access to digital humanities scholarship over time. Most importantly, there was one result that, if confirmed, could have a substantial impact on the design of data curation education programs. In the humanities, center directors and managers appear to resist the

notion that a particular staff role, that of data curator, is specifically needed, preferring instead to develop distributed expertise and responsibilities as part of existing staff roles, calling on institutional resources as needed. This suggests, among other things, that the standing recommendation to place data curation professionals "upstream" in projects may need to be re-envisioned in this context.

2. Introduction

Data curation has been described as "the active and ongoing management of data throughout its entire lifecycle of interest and usefulness to scholarship" (Cragin et al., 2007). Curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time. Originally conceptualized as an e-Science problem precipitated by large amounts of data in digital formats, data curation is an emerging problem for the humanities as well, as both data and analytical practices become increasingly digital (Renear et al., 2009; Crane, Babeu, & Bamman, 2007).

GSLIS received a grant from the Institute of Museum and Library Services (IMLS) to extend the existing data curation specialization within the school's ALA-accredited master's program to include humanities data as well as science data (Renear et al., 2009). Among the activities carried out under the grant was a study of data management and curation practices in the digital humanities. The study was designed and directed by social science researchers from the Center for Informatics Research in Science and Scholarship (CIRSS). The main goals were to assess the levels and types of data curation expertise needed by researchers actively engaged in digital humanities projects and to better understand the potential roles that information professionals trained to meet the unique challenges of working with humanities data in digital formats might fulfill.

3. Methods

To develop a rich picture of data curation practices in the humanities, we employed a case study method, taking digital humanities centers as our case. We chose to focus on established digital humanities centers in preference to libraries, repositories, or individual research teams or scholars for a number of reasons. First, many significant early projects were likely to be located at or affiliated with centers—meaning these centers have experience handling data over longer time scales (Daigle, 2005). Second, centers bring together faculty and staff, and we believe this makes them sites where the most sophisticated curation practices are likely to be found (Zorich, 2008).

We interviewed directors and upper-level staff members at 14 digital humanities centers located in the United States, the United Kingdom, Europe, and Australia with one interview per site. High profile, established centers were chosen that were available and willing to participate in the research. Most centers chosen were located at large research universities but the size of the centers ranged from several staff members to large units. Our intuition was that researchers working at the level of a single project might see data curation as something inextricably tied to their specific job and thus not be able to envision it as a stand-alone function; therefore, upper-level staff with responsibilities for hiring and coordination between projects were chosen in order to elicit views of the curator position from that overarching perspective. Participants were asked about a range of topics related to data management including formats and standards, data storage, security and redundancy, staff roles and background, and significant unsolved problems.

Study participants completed a pre-interview worksheet and their responses were used to guide and focus semi-structured interviews. The pre-interview worksheet also included a series of questions asking participants to rank various categories of skills on a Likert scale ranging from "very important" to "not important at all" for curating humanities data. From the ranked list of skills we were able to develop an overview of researchers' views of data curation in the context of multi-project digital humanities centers. From the interviews, we were able to capture more in-depth information such as complex discussions of tradeoffs and rationales that could not be adequately represented in a simple survey. We are therefore able to report both quantitative and qualitative results from our case study.

4. Results

4.1. Variability and Convergence of Skills

When asked to rank the importance of various areas of expertise needed for a data management professional to be effective working with humanities data, participants revealed an unexpectedly high degree of variability in their answers.

From among a list of 30 kinds of expertise provided in the pre-interview worksheet, at least one study participant gave each category of knowledge the highest score, indicating it was "very important." While

we can rank order the areas of expertise according to an average score, the differences between rankings are not statistically significant by Chi-squared analysis. However, we observe that a handful of skills were both highly ranked on average and showed higher positive skew: every respondent ranked expertise in areas such as interoperability, markup, database design, and metadata as being of moderate importance or higher. Project management also had a high rank order. One surprising result among our findings was the strong emphasis on skills related to teaching and training. This may be due to staff at digital humanities centers being tasked as consultants to scholarly projects or it may simply be due to the expectation within the community that skills will be developed through peer-to-peer training in the course of carrying out job duties. Overall, our results coupled with interview data could not identify a consensus as to the most relevant areas of expertise needed by staff engaged in humanities data curation.

4.2. Organizational and Management Trends

Our interviews reveal a picture of current practice in which the work of data management and curation at digital humanities centers is parceled out among multiple staff members at multiple levels in the organizational hierarchy. Important data curation tasks may be left for scholars or managers of projects to decide individually, or they may be handled by staff, who work on multiple projects for a center, or they may be outsourced to other campus units above the level of the center.

The staff who did have responsibility for data management and curation at the centers we studied were often either those with programming, systems administration, or other IT training, or were people who had received advanced training in a humanities discipline and had taught themselves technical skills.

In keeping with the emphasis on interoperability noted in our quantitative results and also perhaps in response to a changing funding environment and newly available services, we observed a trend in which efforts were being made to move data management expertise from the staff member who had developed it in the course of his or her duties to a center-wide or perhaps institution-wide level where it would be a part of documentation and institutional memory rather than personal memory.

However, our interviews with managers also suggest that even though data management and curation is beginning to be elevated to a higher position in organizations, there is skepticism about the potential role for a data curator at digital humanities centers. Participants in our study were interested in adding skills relevant to data curation to their organizations but rather than doing so in the form of a dedicated position for an information professional, they appeared to be looking for staff with computing or disciplinary skills who also had some training in data curation.

This finding is consistent with another trend we observed. Just as digital humanities centers are already using outside groups such as campus IT or vendors for certain aspects of data management, we noted an increasing orientation to and interest in working with campus-wide services such as institutional repositories to curate humanities data.

5. Discussion

Since effective curation, management, and preservation of data in digital formats involves intervention at every stage of the data lifecycle from creation onwards, it has been a common belief in the data curation community that information professionals trained in curation will need to work “upstream” in scientific labs and digital humanities centers (Swan & Brown, 2008). The current resistance of directors of humanities data centers to such dedicated data curation staff must be taken seriously as it undoubtedly reflects relevant experience and judgment, and their sense of the sorts of arrangements that are likely to succeed. However, our case study in combination with prior work on the role of information work in scientific research suggests that models of provisioning data curation expertise may need to be more nuanced.

As the humanities become increasingly “data-rich,” information science research on data management in the natural sciences becomes increasingly relevant (Choudhury & Stinson, 2007; Renear, Muñoz, & Trainor, 2010). For example, intensive case studies in neuroscience suggest that information services for researchers are likely to be most effective at project stages when information work is most routine or when it is highly speculative, as is often the case with new interdisciplinary research questions or in emerging collaborations (Palmer, 2006; Palmer, Cragin, & Hogan 2007). In the humanities we have also seen that conceptions of what constitutes information or support work and what constitutes professional work within disciplines change in response to the introduction of digital methodologies (Flanders, 2005; Bradley, 2008; McCarty, 2009). While the distribution of curation activities may not follow the same types of (re)arrangements we are seeing in the sciences, we still believe that some data curation work will be

most effective upstream and integrated into scholars' research endeavors, such as at decision points about project planning and re-use value.

As digital curation practices evolve, libraries and institutional repositories will likely take on a larger role in curating humanities data in the future. The results of our study can serve as a useful point of comparison for future work in this area. In addition to having someone who owns data curation problems and manages solutions on a research-center-level, institutions may explore both how to provide services from a central organization (such as the university library) and also, ways to increase the formal, in-service training available to researchers in the digital humanities.

6. Acknowledgments

This work was funded by a grant from the Institute of Museum and Library Services (RE-05-08-0062-08). We have benefited from the expertise of Melissa Cragin, Carole Palmer, and other staff from the Center for Informatics Research in Science and Scholarship in designing and carrying out this research.

References

Bradley, J. (2008). 'What the Developer Saw: An Outsider's View of Annotation, Interpretation and Scholarship'. *Digital Studies / Le champ numérique.* 1(1). http://www.digitalstudies.org/ojs/index.php/digital_studies/article/viewArticle/143.

Choudhury, G.S., & Stinson, T., 2007. 'The virtual observatory and the Roman de la rose: Unexpected relationships and the collaborative imperative'. *Academic Commons.* http://www.academiccommons.org/commons/essay/VO-and-roman-de-la-rose-collaborative-imperative.

Cragin, M. H., Heidorn, P. B., Palmer, C. L., & Smith, L. C. (2007). 'An Educational Program on Data Curation'. *American Library Association Conference, Science and Technology Section.* Washington, D.C., June 25, 2007. http://hdl.handle.net/2142/3493.

Cragin, M. H., Palmer, C. L., Varvel, V., Collie, A., & Dolan, M. (2009). 'Analyzing Data Curation Job Descriptions'. *5th International Digital Curation Conference.* London, U.K., December 2-4, 2009. http://hdl.handle.net/2142/14544.

Crane, G., Babeu, A. & Bamman, D. (2007). 'eScience and the Humanities'. *International Journal on Digital Libraries.* 7: 117-122.

Daigle, B. J. (2005). 'How Do We Sustain Digital Scholarship?'. *Free Culture and the Digital Library Symposium Proceedings.* Martin Halbert (ed.). Metascholar Initiative, Atlanta, GA, October 14, 2005.

Flanders, J. (2005). 'Detailism, Digital Texts, and the Problem of Pedantry'. *Text Technology.* 14(2): 41-70.

McCarty, W. (2009). 'Literary enquiry and experimental method: What has happened? What might?'. *Storia della Scienza e Linguistica Computazionale: Sconfinamenti Possibili.* Liborio Dibattista (ed.). Milan: Francoangeli, pp. 32-54.

Nowviskie, B., & Porter, D. (2010). 'The Graceful Degradation Survey: Managing Digital Humanities Projects Through Times of Transition and Decline'. *Digital Humanities.* London, U.K., July 7-10, 2010. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-722.html.

Palmer, C. L. (2006). 'Weak Information Work and 'Doable' Problems in Interdisciplinary Science'. *Proceedings of the American Society for Information Science and Technology.* 43(1): 1-16.

Palmer, C. L., Cragin, M. H., & Hogan, T. P. (2007). 'Weak information work in scientific discovery'. *Information Processing and Management.* 43: 808-820.

Palmer, C. L., Renear, A. H., & Cragin, M. H. (2008). 'Purposeful Curation: Research and Education for a Future with Working Data'. *4th International Digital Curation Conference.* Edinburgh, Scotland, December 1-3, 2008. http://hdl.handle.net/2142/9764.

Renear, A. H., Dolan, M., Trainor, K., & Muñoz, T. (2010). 'Extending an LIS Data Curation Curriculum to the Humanities: Selected Activities and Observations'. *iSchools Conference.* Champaign-Urbana, IL, February 3-6, 2010. http://hdl.handle.net/2142/15061.

Renear, A. H., Muñoz, T., & Trainor, K. (2010). 'Data Curation Education for the Humanities: Principles & Challenges'. *5th Annual Chicago Colloquium on Digital Humanities and Computer Science.* Northwestern University, Evanston, IL, November 21-22, 2010. http://hdl.handle.net/2142/17421.

Renear, A. H., Teffeau, L. C., Hswe, P., Dolan, M., Palmer, C. L., Cragin, M. H., & Unsworth, J. (2009). 'Extending an LIS Data Curation Curriculum to Include Humanities Data'. *DigCCurr Conference.* Chapel Hill,

N.C., April 1-3, 2009. http://hdl.handle.net/2142/14548.

Smith, A. (2003). *New-Model Scholarship: How Will It Survive?.* Washington, D.C.: Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub114/contents.html.

Swan, A. & Brown, S. (2008). *The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs.* JISC. http://www.jisc.ac.uk/publications/reports/2008/dataskillscareersfinalreport.aspx.

Zorich, D. (2008). *A Survey of Digital Humanities Centers in the United States.* Washington, D.C.: Council on Library and Information Resources. http://www.clir.org/pubs/abstract/pub143abst.html.

When to Ask for Help: Evaluating Projects for Crowdsourcing

Organisciak, Peter
organis2@illinois.edu
University of Illinois, United States of America

A growing online phenomenon is that of crowdsourcing, where groups of disparate people, connected through technology, contribute to a common product. It refers to the collaborative possibilities of a communications medium as flexible and as populated as the Internet. If many hands make light work, crowdsourcing websites show how light the work can be, breaking tasks into hundreds of pieces for hundreds of hands. Building from the growing body of research in the area including the author's work on crowd motivations, this paper outlines the necessary steps and considerations in enriching projects through crowdsourcing.

Though not new, crowdsourcing as it exists online has been enabled by emerging technologies. It has grown out of increasingly efficient – and affordable – forms of communication. Since such collaboration has expanded so quickly, there have been few investigations into the design of crowdsourcing. At the same time, the most successful projects have emerged in an organic nature that many deliberate attempts have failed to replicate, suggesting the need for more investigation in the area. Jeff Howe, who first defined the term and popularized the trend, has explained that "we know crowdsourcing exists because we've observed it in the wild. However, it's proven difficult to breed in captivity" (2008).

The gaps in knowledge of online crowds are quickly being filled however, allowing projects to move away from reliance on serendipity. This presentation derives from recently completed thesis work on the motivations of crowds within crowdsourcing (Organisciak 2010). While it will reflect that study's findings on how, its primary focus is on the equally important questions of why and when in light of those findings. For which tasks is crowdsourcing an appealing option and what resources should be present for a project to adequately motivate the users? A bottom-up classification of crowdsourcing categories is proposed, followed by a checklist of needs that an institution must consider before attempting their own crowdsourcing.

In this study, a sample of 300 crowdsourcing sites was examined and classified. Synthesizing these classifications resulted in a proposed list of eleven non-exclusive categories for crowdsourcing, six describing method and five describing structure. Methods include encoding, creation, idea exchange, skills aggregation, knowledge aggregation, and opinion aggregation. Additionally, there are financial, platform, gaming, group empowerment, and ludic structures observed within these systems. Derived from existing systems, these categories and their variants offer unique design patterns and best practice cases that can assist in assessing the types of tasks at which they excel.

Appropriateness of the task is just one facet of running a crowdsourcing project. The other consideration is whether a project offers a return that potential participants would find rewarding. In addressing this, a content analysis was used to identify site design mechanics related to user experience in thirteen cases spanning the breadth of the identified categories. These mechanics were then discussed in a series of user interviews to determine what users truly care about. In this study, a collection of primary and secondary motivators are proposed as foundational considerations in running a project. The primary motivators seen in the user interviews were interest in the topic, ease of entry and of participation, altruism and meaningful contribution, sincerity, and appeal to knowledge. A final one, financial incentive, is perhaps the most blunt. Secondary motivators include indicators of progress and reputation (i.e. “cred”), utility, fun, system feedback, social networking, and fixed windows (i.e. well-groomed quality).

An understanding of the nature of crowdsourcing holds notable benefits to scholarship in the humanities and social sciences. Most significantly, this is because it allows large-scale insights into the qualitative and the abstract, those areas inextricably linked to the limits of manpower, unable to be delegated to computing power. “What is the sentiment of this sentence”, is the type of question a crowdsourcing site may ask (Mechanical Turk, May 2nd 2010), if not always as directly. Since much work in the arts cannot easily be quantified, logistics and resources often limit humanities research to a balance between breadth and depth; crowdsourcing offers an escape from this issue.

Consider one task that is often seen in existing crowdsourcing sites: crowd-encoded classification. Classification tasks are dependent on the person-hours available, because person-hours are the only dependable way to approach these tasks. Whether directly or incidentally, online crowds can effectively encode or classify content. Though the reliability of the end product is often far below that of a professional encoder, large-scale crowd projects can often account for this through multiple independent classifications, measuring consistency and reliability through agreement. Galaxy Zoo, an effort from Oxford to classify galaxies, found crowdsourced data to be within 10% agreement with the same data classified professionally (Lintott et al. 2009). The high quality of work is especially notable because the experiment and its follow-ups received their 60 millionth classification in April 2010.

Flickr Commons, an initiative to put photo archives on a photo-sharing community, is a similar project that – by way of community-based research, information and tagging – has enriched the metadata of hundreds of Library of Congress photographs in the United States of America (Springer, et al. 2008). Another pilot project involving public tagging, by the National Library of Australia, concluded that “tagging is a good thing, users want it, and it adds more information to data. It costs little to nothing and is relatively easy to implement; therefore, more libraries and archives should just implement it across their entire collections” (Holley 2010). The National Library of Australia followed through on this recommendation.

Such projects are often greeted with suspicion in professional or scholarly communities. The National Library of Australia report notes that "institutions who have not implemented user tagging generally perceive many potential problems that institutions who have implemented user tagging do not report" (Clayton et al. 2008 qtd. in Holley 2010). The Library of Congress report similarly notes many concerns that critics provided, such as: “Would fan mail, false memories, fake facts, and uncivil discourse obscure knowledge? … Would the Library lose control of its collections? Would library catalogs and catalogers become obsolete?...Would history be dumbed-down? Would photographs be disrespected or exploited?” (Springer et al. 2008). In both cases, the reports state that the concerns, within the respective project’s experiences, have not manifested.

Encoding is a notable use of crowdsourcing in academia, but not the only one. Some projects, such as the Suda On Line, benefit from collected contributions of expertise and knowledge. Suda On Line is a project to translate a Byzantine encyclopedia, Suda, into English for the first time. It has been steadily progressing since 1998, producing a comprehensive resource while staying at a manageable participation scale (Mahoney 2009). In other cases, crowdsourcing allows public and volunteer projects to compete with the scale and quality of commercial projects, as has

been seen in OpenStreetMap, Project Gutenberg, and many open source projects.

As crowdsourcing continues to be tested – and if it continues to be successful – in public institutions, understanding how to undertake such projects will become more important. The benefits are being stated, and the scale and openness on which public institutions operate makes them a compatible beneficiary of crowdsourcing activities. Users appear especially altruistic toward public projects, emphasizing in this study their preference for meaningful engagement with institutional workings over symbolic outreach.

The study informing this work is large, and my hope is to provide a digestible account of its results. The reason for this goal is straightforward: there is still much work to be done in understanding the mechanics of crowdsourcing, but the potential is great. I hope that the sharing of this foundational work will encourage others to explore further.

2. Acknowledgements

This study owes a great debt to Lisa M Given, my thesis advisor, as well as additional committee members Geoffrey Rockwell and Stan Ruecker.

References

Holly, Rosé (2010). 'Tagging Full Text Searchable Articles: An Overview of Social Tagging Activity in Historic Australian Newspapers August 2008 — August 2009'. *D-Lib.* 12(1/2). http://www.dlib.org/dlib/january10/holley/01holley.html#4.

Howe, Jeff (2008). *Crowdsourcing.* .

Lintott, Chris, et al (2010). 'Galaxy Zoo : Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey'. *arXiv.* 0804.4483. http://uk.arxiv.org/abs/0804.4483.

Organisciak, Piotr (2010). *Why bother? Examining the motivations of users in large-scale crowd-powered online initiatives.* http://repository.library.ualberta.ca/dspace/handle/10048/1370.

Springer, Michelle, et. al. (2008). *For the Common Good: The Library of Congress Flickr Pilot Project.* .

Mahoney, Anne (2009). 'Tachypaedia Byzantina: The Suda On Line as Collaborative Encyclopedia'. *Digital Humanities Quarterly.* 3.1.

The Cultural Impact of New Media on American Literary Writing: Refining a Conceptual Framework

Paling, Stephen
paling@wisc.edu
School of Library and Information Studies, University of Wisconsin-Madison

1. Introduction

This paper describes a survey study [1] that is part of an ongoing effort (Paling & Martin, 2009; Paling, 2008; Paling & Nilan, 2006) to extend Social Informatics(Kling, 1999) to the study of literature and art. This series of studies has focused on the emergence of new forms of literary expression offered by information technology, and whether and how those possibilities are finding a place alongside traditional forms of expression in American literary writing. The study is meant to be complementary to more discursive, hermeneutic views of literary work. Discursive discussion of literary work provides rich descriptions of work by selected authors. In contrast, a survey enables us to look at the actions of literary community members in the aggregate. This study is based, in part, on the idea that various forms of inquiry into emerging literary practices, taken together, will provide a more complete picture than any one form of inquiry alone can provide. Different approaches to inquiry need not be seen as oppositional.

2. The Previous Studies

The previous studies in this series (Paling and Martin, 2009; Paling, 2008; Paling & Nilan, 2006) developed, and found support for, a conceptual framework made up of four key values that typify American literary writing:

1. Positive Regard for Symbolic Capital.
2. Negative Regard for Immediate Financial Gain.
3. Positive Regard for Autonomy.
4. Positive Regard for Avant-garde-ism.

Positive regard was operationalized as responses that indicated admiration or desire for particular qualities, e.g., a desire to read an author's work based

on previous work by that author. Similarly, *avant-garde* was operationalized as valuation of particular characteristics, e.g., a preference for literary work that is fresh or innovative, or electronic work that has characteristics that cannot be produced in print form.

The previous studies examined the ways in which literary authors could use information technology to change those key values. They posited, and found support for, the idea of *Intensifying Use of Technology*, which has three characteristics:

1. Recognition of new forms of support for a value.
2. Incomplete rejection of traditional forms of support for the value.
3. Placement of greater emphasis on the newer forms of support.

For example, a fiction writer could use a hypermedia editor to produce non-linear, electronic fiction (recognition). She might prefer this newer form of fiction (greater emphasis on newer form of support), but also continue to write more traditional work (incomplete rejection).

The original study in the series (Paling & Nilan, 2006) involved interviews with a purposive sample for heterogeneity (n=36) of editors of American little magazines. That study used primarily qualitative methods. The second study (Paling, 2008) involved the same sampling method (n=22), but focused on American literary authors. Paling (2008) used both qualitative data as well as quantitative data derived from Likert scales. The third study in the series (Paling & Martin, 2009) was a pilot survey with a random sample (n=84), and that study led directly to the development of the current study. The original studies all showed the presence of intensifying use of technology, but differed in terms of how common that phenomenon seems to be. Because two of the three samples were non-random, and all three were relatively small, a larger, random sample will lead to firmer conclusions about the research questions.

3. The Current Study

The current study concentrated on two research questions:

1. RQ1: Do members of the American literary community show support for Positive Regard for Avant-garde-ism?
2. RQ2: Do the actions of members of the American literary community reflect Intensifying Use of Technology with regard to Positive Regard for Avant-garde-ism?

Positive Regard for Avant-garde-ism was selected from the four key values that make up the conceptual framework because it is the most relevant to the use of information technology. It is directly relevant, for example, to how much participants value the use of information technology to produce innovative writing that cannot be done in print. However, the other key values play an important role in establishing the context within which a value such as Positive Regard for Avant-garde-ism comes into play, and they were retained as an important part of this study.

The current study represents a clear methodological progression along the line of research begun in the earlier studies. A total of 900 invitations for participation were sent out. The names were selected randomly from a sampling frame built based on publicly available lists from The Association of Writers and Writing Programs and the Modern Language Association, as well as print directories such the Council of Literary Magazines and Small Presses' *Literary Press and Magazine Directory*. This yielded a sample of exactly 400 participants. All of the potential participants live, work, or study in America, or work for a publisher whose primary presence is in America. An international sample would be desirable in the future, but was beyond the scope and funding level of this study.

The respondents were asked to complete a mail survey composed of brief yes/no or checklist questions, as well as questions that included Likert-type numeric scales. The questions were refined versions from the interviews in the earlier studies in the series. Data, and respondent comments, from the earlier studies uncovered minor problems with question wording, scaling, etc. Because of that, the data from this study cannot be combined with data from the earlier studies. Conclusions from the current study, though, should be given greater weight than conclusions from the earlier studies because of the larger, random sample and the refined questions.

4. Reconciling the Studies

Much of the apparent difference between the previous studies seems to have resulted from the different sampling methods used in the studies in this series. The first two studies (Paling, 2008; Paling & Nilan, 2006), as mentioned earlier, used purposive sampling for heterogeneity. In other words, an effort was made to seek out editors and authors who were actively involved in producing literature with a substantial electronic component. The strength of purposive

sampling is that it allows this kind of effort to closely examine different segments of a community such as people who participate in American literary writing. The weakness of such sampling, though, is that it is very difficult to create a purposive sample that reflects not just the presence of particular phenomena, but also an accurate sampling of how widespread the phenomenon is.

The pilot survey (Paling & Martin, 2009) did find limited demonstration of Intensifying Use of Technology. One respondent showed unambiguous evidence of Intensifying Use of Technology. A number of other respondents (8) demonstrated somewhat similar, but less pronounced, patterns of Intensifying Use of Technology. This would suggest that somewhere around 10% of the American literary community demonstrates Intensifying Use of Technology, although the conceptualization of that phenomenon may need to be altered to reflect the opinions of community members who place strong positive value on technological innovation in literature, but who do not actually place greater emphasis on such forms of literature.

Taken together, the previous studies suggest that we need to modify, but not reject, the concept of Intensifying use of Technology. Intensifying Use of Technology is a real phenomenon, but is not, to date, widespread in American literary writing. More importantly, the American literary establishment demonstrates very limited levels of the phenomenon. This conclusion is very similar to the conclusion reached by Rettberg (2009). Rettberg argued that electronic literature constitutes a literary avant-garde, but an avant-garde that is not part of any institutionalized mainstream. However, Rettberg's work represents the analysis of an individual directly involved in those efforts, and did not involve structured data gathering to address the actions of literary community members in the aggregate. The current study takes the contrasting approach, gathering data from respondents across the literary community to begin developing a larger picture of how information technology is influencing contemporary American literary writing. The size and type of the sample used in this study should help resolve any ambiguities raised in the previous studies.

References

Kling, R (1999). 'What is Social Informatics and why does it matter?'. *D-Lib* , vol. 5, no. 1. Retrieved September 12, 2006. Available from . . `http://www.dlib.org/dlib/january99/kling/01kling.html` [Accessed November 15, 2009]. .

Paling, S, Nilan, M (2006). '"Technology, values, and genre change: the case of little magazines"'. *Journal of the American Society for Information Science and Technology*, , vol. 57, no. 7: pp. 862-872, . . .

Paling, S (2008). '"Technology, genres, and value change: literary authors and aesthetic use of information technology"'. *Journal of the American Society for Information Science and Technology*, , vol. 59, no. 8: pp. 1238-1251,. . .

Paling, S, Martin, C (2009). '"Toward a theory of technological transformation in artistic genres"'. *American Society for Information Science and Technology Annual Meeting, Social Informatics Symposium*, paper presented at the , Vancouver, BC. . .

Rettberg, S (2009). '"Communitizing electronic literature"'. *Digital Humanities Quarterly*, , vol. 3, no. 1. Available from . . `http://digitalhumanities.org/dhq/vol/3/2/000046.html` [Accessed November 15, 2009]. .

Notes

1. Data gathering is under way at the time of this writing. The data gathering will be complete by the time of the conference.

Browsing Highly Interconnected Humanities Databases Through Multi-Result Faceted Browsers

Pasin, Michele
michele.pasin@kcl.ac.uk
Department of Digital Humanities, King's College, London, UK

Faceted browsing is a recent paradigm in search interfaces that allows users with little familiarity of a subject domain to quickly explore the contents of databases or other structured data sources. The underlying principle of this approach can be traced back to the work of Indian librarian S.R. Ranganathan, who, in contrast with traditional top-down, taxonomical approaches to subject classification, in 1933 developed a method for organizing subjects in a bottom-up and non-hierarchical fashion. According to this model, a classification system can be created by combining together subject descriptors chosen from a number of non-exclusive and non-hierarchical facets – e.g., in the context of classifying books, these can be genre, date or author. This method supports the generation of a flexible system that better represents the multitude of perspectives we could use to represent knowledge (Broughton, 2004).

As a result of the adoption of these ideas in computer science, researchers have been creating search interfaces that allow the exploration of digital resources through the manipulation of filters describing important features of a subject domain (Broughton 2002). A well-known pioneer in this area is Marti Hearst with her work on the Flamenco faceted browser (Hearst, 2002), followed by a number of similar approaches that, in general, aimed at creating more compelling and easy-to-use user interfaces (Hearst, 2008) (Capra, 2007) or at providing software packages that can work with different types of data sources; these may vary from manually editable JSON files (Huynh, 2007) to databases (Stuckenschmidt, 2004) and RDF triplestores (Oren, 2006). A number of projects have also proven the usefulness of this type of interfaces to the aim of facilitating the navigation of large repositories of humanities data, such as artwork images (Hildebrand, 2006), music resources (Bretherton, 2009) or multi-genre collections (McGann, 2007).

The success of faceted interfaces can be related to the fact that they implement a schema-less approach to classification, that is to say, they make available to the user a number of co-existing search dimensions that can be simultaneously used to browse and preview the contents of a digital repository. Many are the proven advantages of such an approach (Perugini, 2010): first, users are never asked to 'guess' the right search terms, as it happens in classic keyword search interfaces; second, inconclusive searches are prevented; third, non-experts can easily 'get a feeling' for the significance and meaning of the data available just by looking at the available facets, thus increasing their understanding of the domain. In conclusion, this type of interfaces simplify enormously the exploration of a digital repository, and, using the words of Nowivskie, they make it easy to "explore lateral relationships" to the point that they open "possibilities for algorithmic serendipity in research" (Nowivskie, 2007).

An important feature that most of the existing faceted browsers have in common is that the different facets available equally concur to the selection of a single result-type. For example, by manipulating variables such as the *color* and the *making* of cars, we can navigate a data-space of available cars; by choosing filters representing information about *genres*, *publishers* and *years of publication* we can easily narrow down a result list of books; or, in the context of a prosopographical database, by accumulating descriptors about people's *forenames*, *surnames*, or *gender* we would be able to refine our search for the *individuals* mentioned in the database.

In our work, we intended to push the boundaries of this approach by creating a faceted browsing engine that, given the same set of selectable facets, can be used to search for 'ontologically distant' entity types. For example, in the context of a prosopographical database, by manipulating the same group of filtering options, we aimed at letting users search not just for *people*, but also for *factoids* and *sources* (cf. Figure 1). In doing so, we assumed that rich and highly interconnected humanities databases call for more powerful search mechanisms; such mechanisms should be capable of revealing the intricacies of a subject domain to the casual learner, and, at the same time, of providing a higher level of 'algorithmic serendipity' to the academic scholar. In other words, we aimed at making more visible the large number of search pathways a highly structured database can make possible - as opposed to hiding this complexity by providing a unique entry point to the wealth of data available. This means that, as shown in fig. 1,

by using facets typical of the 'people' result-type (such as *surname* or *gender*) we would like to be able to search for 'sources' or 'factoids'. Or, by choosing facets typical of the 'sources' result-type (such as *document category* or *language*) we may want to filter results when searching for 'people'.

With this vision in mind we created DJFacet[1], a faceted browsing engine that lets users create powerful, multi-result search interfaces. DJFacet is written in Python and is based on Django, a popular web application framework[2] that facilitates the development of database-driven websites by providing functionalities that speed up the creation of repetitive tasks. In particular, one key component of this framework is the Object-Relational Mapper (ORM), that is, a set of functions that provide programmers with a level of abstraction between the database and the application language; as a result, it is possible to invoke complex database queries without having to write any SQL code. This makes the whole application easier to manage and more portable across different database engines.

By building on the functionalities of Django ORM, DJFacet provides a customizable and easy-to-use environment for creating database-driven faceted search applications. The underlying idea of DJFacet is that since a data-structure had already been designed and fine-tuned when the database was created, that same structure could be used to define the search dimensions of the faceted browser with little extra work required. An instance of DJFacet can run just by creating an initialization file in which we define which are the available facets and what 'behaviour' they have with respect to the database schema. The software then calculates automatically all the remaining query-paths needed to show the results in the various searches.

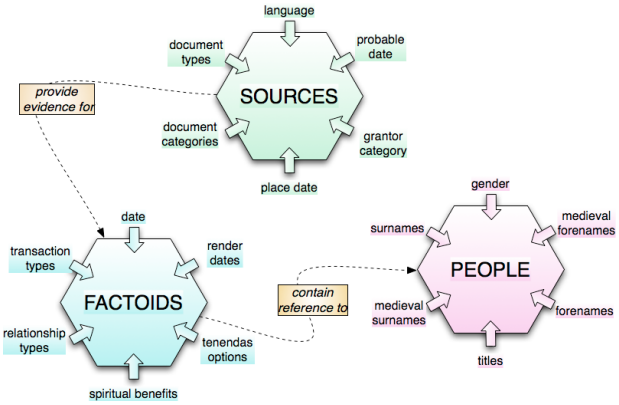
At the time of writing, we tested DJFacet's approach with two humanities databases. The Paradox Of Medieval Scotland project[3] (POMS) investigates how a recognizably modern Scottish identity was formed during the period 1093-1286. Drawing on over 6000 contemporary charters, it provides biographical information about all known people in Scotland during that period. In this context we built a search interface that features 29 facets, organized into 5 groups (cf. figure 2). The result types are 3 (*people*, *factoids* and *sources*).

The Early Modern London Theatres[4] (EMLoT) project provides its users with a major encyclopedic resource on the early London stage, as well as a comprehensive historiographical survey of the field. EMLoT identifies, records and assesses transcriptions from primary-source materials relating to the early London stage, as found in secondary-source print and manuscript documents. In this case the faceted search interface contains 24 facets, organized into 5 groups. The result types are 7 (*transcription records, primary sources, secondary sources, events, people, troupes, venues*).

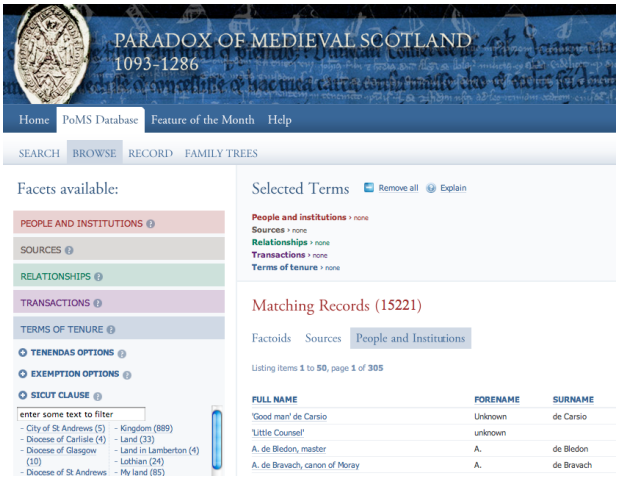
By using DJFacet it was possible to allow the formulation of queries that might not be immediately obvious to the user. For example, in POMS it became trivial to search for Charters mentioning transaction events in which 'beneficiaries' of name 'William' acquire something on the day of the 'Feast of St Patrick '. Also, the search interface provided users with more chances of coming across interesting connections in the available materials. This was made possible by the fact that the facets used in the search are ontologically distant from the respective result-types.

However, despite the fact that this approach proved to be, from the logical and computational point of view, completely feasible, it also opened up a number of research questions from the point of view of the meaning of these multifaceted searches across different results types. In other words, we realized that often the accumulation of filters ontologically distant from each other could be hardly translated by the end user into real-world questions; analogously, the opposite may happen, in so far as simple type of searches may be impeded by the highly structured architecture of a faceted browser.

In order to provide some answers to this issue and lay the ground for a more scientific discussion of the problem we are currently carrying out a user evaluation study with humanities scholars. The purpose of the experiment is to discover the degree to which humanities scholars can make sense of the search mechanisms provided by our faceted browser, and, indirectly, of the complex data structures often necessary for representing humanities subjects. We will report on these findings at the conference, together with a deeper analysis of the implications of using multi-result faceted browsers in the context of complex humanities datasets.



The facets allowing 'entry' to a prosopographical database.



Screenshot of the faceted browser for the POMS database.

References

Bretherton et al. (2009). 'Integrating musicology's heterogeneous data sources for better exploration.'. *10th International Society for Music Information Retrieval Conference.* 2009.

Broughton (2004). *Essential classification.* London: Facet Publishing.

Broughton (2002). 'Faceted classification as a basis for knowledge organization in a digital environment: the bliss bibliographic classification as a model for vocabulary management and the creation of multidimensional knowledge structures'. *The New Review of Hypermedia and Multimedia.* vol. 7 (1): 67-102.

Capra et al. (2007). 'Effects of structure and interaction style on distinct search tasks'. *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries.* 2007, pp. 442-451.

Hearst (2008). 'UIs for Faceted Navigation: Recent Advances and Remaining Open Problems'. *HCIR08 Second Workshop on Human-Computer Interaction and Information Retrieval.* 2008.

Hearst et al. (2002). 'Finding the flow in web site search'. *Communications of the ACM, Special Issue: The consumer side of search.* vol. 45 (9).

Hildebrand et al. (2006). '/facet: A Browser for Heterogeneous Semantic Web Repositories'. *ISWC'06: Proceedings of the International Semantic Web Conference.* 2006, pp. 272-285.

Huynh et al. (2007). 'Exhibit: lightweight structured data publishing'. *WWW '07: Proceedings of the international conference on World Wide Web.* 2007, pp. 737-746.

Oren et al. (2006). 'Extending Faceted Navigation for RDF Data'. *ISWC'06: Proceedings of the International Semantic Web Conference.* 2006, pp. 559-572.

McGann and Nowviskie (2005). *NINES: a federated model for integrating digital scholarhip. White paper*.

Nowviskie (2007). *COLLEX: semantic collections & exhibits for the remixable web. White paper*.

Perugini (2010). 'Supporting multiple paths to objects in information hierarchies: Faceted classification, faceted search, and symbolic links'. *Information Processing & Management.* vol. 46 (1): 22-43.

Stuckenschmidt et al. (2004). 'A topic-based browser for large online resources'. *Engineering Knowledge in the Age of the SemanticWeb, LNCS.* vol. 433-448: 433-448. `http://URL`.

Notes

1. The software is open source and freely available online at the url `http://code.google.com/p/djfacet/`
2. `http://www.djangoproject.com/`
3. `http://www.poms.ac.uk`
4. `http://www.emlot.kcl.ac.uk`

Civil War Washington: An Experiment in Freedom, Integration, and Constraint

Price, Ken
kprice2@unl.edu
University of Nebraska-Lincoln

Barney, Brett
bbarney2@unl.edu
University of Nebraska-Lincoln

Lorang, Liz
liz.lorang@gmail.com

Civil War Washington (CWW) is a thematic research collection that strives to enable users to visualize, analyze and interpret the physical, social, cultural, and political transformation of Washington, D.C. The development of Washington, D.C., during the Civil War is pivotal in American history. When the Compensated Emancipation Act went into effect on April 16, 1862, Washington became the first emancipated city—and the country's largest and most important magnet for freed and runaway slaves. From that moment forward, the city would lead the nation in the sometimes tortuous route from slavery to freedom and from an entrenched system of legal inequality to a new commitment to equality for all. Our work on slavery, race, and emancipation in Washington, D.C., is crucial to our larger long-term study of the city in this time of crisis. We are already studying Civil War Washington from a medical perspective (the number of hospitals jumped from three to nearly one hundred making it a city of hospitals), from a military perspective (the city was the prized objective of Southern military strategy and in response the Union army made it the most fortified city in the world), and in fact from numerous other perspectives as well. With the assistance of a collaborative research grant from the National Endowment for the Humanities, our emphasis in 2010-2013 is to study the history of race, slavery, and emancipation in the city, a story of national importance.

The transformation of the U.S. capital has received surprisingly little sustained examination. One reason for this, no doubt, is that developing a rich and accurate understanding of the city's remaking requires not only access to but also synthesis and analysis of large and diverse sets of data, most of which exist only in analogue form. Our project, for example, draws on government reports, journalism, legal documents, diaries, census records, correspondence, city directories, poems, maps, and photographs. A further (and we believe necessary) complication is added by our desire to make a temporally-aware and user-manipulable GIS a significant constituent of our project, perhaps in some ways even its core.

The number of projects, both emergent and established, that claim an interest in GIS and place-based scholarship seems to grow daily. This developing interest is reflected in an increasing number of seminars and training opportunities focused on GIS for the humanities, including at the University of Victoria's Digital Humanities Summer Institute, the Digital Humanities Observatory Summer School in Dublin, and the NEH-funded Geospatial Institute hosted by Scholars' Lab at the University of Virginia. In addition, there is a growing body of work on the value of spatial analysis in the humanities, and historians have taken a leading role in theorizing and conceptualizing the integration of GIS into humanities research, as well as in advancing arguments informed by GIS-enabled research.

In recent years, at conferences such as this, a fair number of projects have been presented as models for addressing the still-daunting set of obstacles to the use of GIS for humanities projects.[1] Our experience on Civil War Washington for the past five years, though, suggests that several important challenges have not yet been adequately investigated (let alone addressed). Our talk, then, is designed to be not a celebration of goals achieved but a case study for the consideration of several large issues that face our project and others like it.) Whereas most projects presented as models began with fairly well defined sets of data,[2] ours began with a research question to which we want to apply as comprehensive a variety of data as possible. Our goal is not the digitization of materials for the sake of digitization but the exploration of a complex set of questions about the transformation of Washington, D.C. Further, our goal ultimately is to produce both a scholarly argument of our own and a resource that will enable our users to perform original, and truly meaningful, research based on our data and interfaces. Given these aims, on what basis should one decide the best technologies and methods for data capture, storage, and retrieval? We have responded to these questions based in part on the expertise of project members and the technologies that are familiar to us and local support staff. This strategy has the benefit of quicker development and an ever-more-refined knowledge in specific tools and

technologies. Too heavy a reliance on the technologies and methods that one already knows, however, can lead to overlooking a more appropriate approach or set of approaches. 2) Given the fact that capture and storage of geo-referenced textual data can be accomplished in several different ways (e.g., through wholly TEI-XML, through assigning atomistic textual units to database fields, or through a combination strategy that uses both XML and database), what principles should guide the adoption of a particular encoding strategy? What tools exist or are most easily imaginable for making such data available for a wide variety of research approaches? TEI began providing for the georeferencing of textual data only with the release of P5 in late 2007, and we believe that the integration of textual and geospatial data remains an underdeveloped area to which the DH community should give greater thought. Surely it has potential beyond its rather modest application in existing projects. 3) How should a project deal with the catch-22 situation of wanting to develop and enable geospatial and historical analysis using open source tools even as the most adequate tools for the task, regrettably, are proprietary and the use of these proprietary tools is taught in leading DH institutions? In other environments, work on open source GIS has emerged as part of the larger open source and open access movement. The Open Source Geospatial Foundation, for example, is working to "promote the collaborative development of open geospatial technologies and data."[3] Humanities scholars, however, do not seem to be involved in the organization. What is and should be the role of humanities scholars in the developer community for GIS software? Does the existing open source GIS software originally conceived of by organizations such as the Army Corps of Engineers (Grass GIS) or NASA (MapServer) meet the needs of humanities scholars? We have considered these questions not merely in the realm of abstract ideals, but as immediate and pressing concerns; we wish to use our project's responses to them not as recommendations but as opportunities for critical reflection and discussion.

Notes

1. Martyn Jessop, in his talk At the first Digital Humanities conference in 2006, "The Inhibition of Geographical Information in Digital Humanities Scholarship," presented a list of barriers to the application of GIS to humanities projects. This talk, later published in Literary and Linguistic Computing (April 2008), is largely still relevant. In addition, see also the report of the 7th Scholarly Communication Institute at the University of Virginia, "Spatial Technologies and the Humanities": http://www.uvasci.org/wp-content/uploads/2009/10/sci7-published-full1.pdf
2. For example, see the following DH2010 abstracts: Ian Gregory, "GIS, Texts and Images: New Approaches to Landscape Appreciation in the Lake District," Elton Barker, et al., "Mapping the World of an Ancient Greek Historian: The HESTIA Project," and " Wayne Graham, "A New Spatial Analysis of the Early Chesapeake Architecture," available at http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/
3. See http://www.osgeo.org/content/foundation/about.html

A Data Model for Visualising Textuality – The Würzburg Saint Matthew

Rehbein, Malte
malte.rehbein@uni-wuerzburg.de
Würzburg University

1. Abstract

This short paper presents the ongoing work and the considerations behind the project "Visualising Textuality – New Interfaces to Historical Texts".

The project, supported by the EU FP7 Marie Curie Scheme, aims at implementing a "knowledge environment" (Siemens et al) to explore and understand better early medieval textual practices and pre-scholastic Christian scholarship. The project starts from a manuscript from the University Library of Würzburg (M.p.th.f.61), dated back to the second half of the 8th century AD and most likely of Irish provenance. This is a parchment manuscript with 34 leaves containing a text of Matthew's gospel along with extensive interlinear glossing and 30 cedulae (parchment slips) containing commentary material bound between the pages (Fig. 1).

Fig. 1: Excerpt from M.p.th.f.61 with comments on parchment slips.

It is also an intriguing object of historical studies: its numerous glosses (interlinear and marginal) as well as commentaries (on parchment slips) allow insight into practices of compilation and use of this manuscript (cf. Fig. 1). However, it poses many challenges to editors and researchers: different layers of writing (strata) have been identified; the arrangement (mise-en-page) of glosses and commentaries in relation to the Gospel text and to each other conveys important information but is not easy to follow; the "intertexts" cross the logical segments of the text and the physical boundaries and dimensions of the document pages; the texts themselves, especially the commentaries, have their own history of transmission; and they recite, vary from and refer to other commentaries such as Eusebius, Jerome or Isidor, who also refer to other commentaries and biblical texts etc. Cahill summarizes that the Würzburg Matthew is "a complicated jumble and not a tidy bundle" and requires further research (2002:25). His statement applies to all levels of investigation: the physical (document), the logical (texts and contexts) and the level of space and time (cf. Fig. 2).

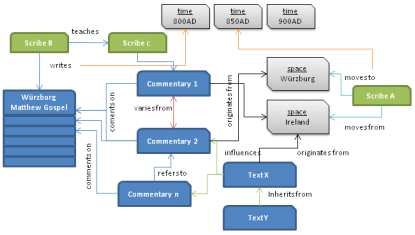


Fig. 2: Model illustration of the complex network of relations based on the Würzburg glosses (not complete).

It is the aim of this research project to find adequate means for representing such a complicated network of information, for visualising these relations and for allowing the researcher to navigate through this data in form of a "knowledge environment". The fact that this all is a "complicated jumble" has to be seriously taken into consideration. The data that is outlined here consists of numerous combinations – of semantic relations, chronological dependencies, spatial transmissions, topological information – each of which may carry an important detail for research. What serves for the first time for an in-depth and comprehensive research, has its risk in getting lost in information. Thus, it seems to be crucial for the scholar to focus his or her attention up on what (s)he is interested in. Support for this needs to be provided by the knowledge environment.

The project is in the first year of its three years implementation plan. At the time of the DH conference (June 2011), most of the textual work on the manuscript shall be completed, so that preliminary

considerations about the setup of the envisaged knowledge environment and what is more important about the underlying data model will be discussed and presented. This data model is an abstraction of document, text and context and encompasses entities of different types (such as segments of texts, zones in a document, time, space, scribes, sources, metadata) and relations among them (such as “origins from”, “uses”, “was written by”) as well the interface to external data (such as the *Patrologia latina* database) to link to text variants and contexts.

References

Bischoff, B. (1954). 'Wendepunkte in der Geschichte der lateinischen Exegese im Frühmittelalter'. *Sacris erudiri.* .

Cahill, M. (2002). 'The Würzburg Matthew: Status Quaestionis'. *Peritia.* .

Card, S., J. MacKinlay, B. Shneiderman. (1999). *Readings in information visualization: Using vision to think.* .

Chen, C. (2007). *Information Visualization: Beyond the Horizon.* .

Dahlström, M.. *The Compleat Edition Text editing, print and the digital world.* Pp. 27–44.

Eggert, P. (2005). 'Text-encoding, Theories of the Text, and the Work-Site'. *Literary and Linguistic Computing.* 20.4: 425-35.

Flanders, J. (2009). 'The Productive Unease of 21st-century Digital Scholarship'. *Digital Humanities Quarterly.* 3.3.

Gabler, H. (2007). 'The Primacy of the Document in Editing'. *Ecdotica.* , pp. 197-207.

Gervers, M. (2007). 'New Methods for the Analysis of Digitized Medieval Latin Charters'. *Historisches Forum.* 10: 482-500.

Gippert, J. (2002). *The Old Irish "Würzburg" Glosses.* `http://titus.fkidg1.uni-frankfurt.de/texte/celtica/wbgl/wbgl.htm`.

Gwynn, A. (1952). 'The Continuity of the Irish Tradition at Wurzburg'. *Herbipolis jubilans.* , pp. 57-81.

Kelly, J. (1993). 'The Würzburg Saint Matthew'. *Würzburger Diözesangeschichtsblätter.* 55: 5-12.

Köberlin, K. (1891). *Eine Würzburger Evangelienhandschrift.* .

Ó Cróinín, D. (1982). 'Mo-Sinnu Moccu Min and the Computus of Bangor'. *Peritia.* 1: 281-95.

Rehbein, M. (1999). 'Komplexe Textkritik in dynamischer Darstellung: Ein Modell für digitale Texteditionen'. *HSR.* 24.1: 113-44.

Robinson, P. (2009). 'The Ends of Editing'. *Digital Humanities Quarterly.* 3.3.

Schepss, Georg. (1887). *Die ältesten Evangelienhandschriften der Würzburger Universitätsbibliothek.* .

Shillingsburg, P. (2006). *From Gutenberg to Google: Electronic representations of literary texts.* .

Siemens, R. et al. (2009). 'Toward a Conceptual and Theoretical Foundation for New Research on Books and Knowledge Environments'. *Digital Studies.* 1.2.

Stokes, W.. 'The Old-Irish glosses at Würzburg and Carlsruhe'. *London: Philological Soc* , 1887..

Sutherland, K.. 'Being Critical: Paper-based Editing and the Digital Environment'. *Text editing, print and the digital world.* Pp. 13–25.

Teeuwen, M.. 'The Impossible Task of Editing a Ninth-Century Commentary: The Case of Martianus Capella'. *Variants.* 6.

Thurneysen R. (1901). 'Das Alter der Würzburger Glossen'. *Zeitschrift für Celtische Philologie.* 3.1: 47-54.

Tufte, E. (2008). *Envisioning information.* .

Toward a Demography of Literary Forms: Building on Moretti's Graphs

Riddell, Allen B.
allen.riddell@duke.edu
Duke University

Why do novelistic genres end? Why do we see gothic and industrial novels, *Bildungsromane* and mysteries, all disappear after periods of popularity during the 18th and 19th centuries? How literary forms—novelistic genres in particular—come and cease to be has long been an area of inquiry, and the work of Franco Moretti (2005) in *Graphs, Maps, Trees* has given the topic new energy.

The years since the publication of "Graphs" in 2003 have seen the resources available for investigating patterns in 19th century literary production expand immeasurably. For example, scans of over 7,800 volumes of 19th century British novels are now available from the University of Illinois-Urbana-Champaign's collection alone. Those interested in extending or challenging Moretti's observation that novelistic genres tend to arrive in "bursts" linked to social generations—or, in general, in contributing to Moretti's proposed "sociology of literary form"—now have access to a wealth of new data. Given the continuing interest that Moretti's work has generated among students of the human, social, and natural sciences, this represents an important opportunity.

My contribution explores further the prospects for a "demography of literary forms," building on Moretti's proposal for research at the intersection of literary history and sociology.[1] First, I consider opportunities to improve Moretti's original "generational model" of cycles in literary forms, offering new methods to remedy identified evidential gaps. For example, Moretti's periodizations of genres—e.g. Courtship Novel, 1740-1820—have been criticized as too neatly falling on certain "focal dates" such as years falling at the end of a decade (Shalizi, 2006). I present new evidence in support of this criticism and demonstrate a method for making periodizations reproducible by others. Being able to reproduce Moretti's results will hopefully make the research program itself more open to experimentation and invite collaboration. My method uses bibliographic databases to establish the period during which the vast majority (~90%) of the novels in a given genre were published. This provides a reproducible periodization suited to an inquiry into social history.[2] I also briefly discuss the application of classification algorithms from machine learning to identify possibly overlooked genres in Moretti's dataset. The new method for periodization is demonstrated in detail for two of the forty-four genres in Moretti's dataset, the silver fork and Newgate novels.[3]

Second, in order to offer an alternative to the generational model, I argue for and attempt to test the hypothesis that the observed clustering of genre appearances and disappearances can also be explained by positing a "carrying capacity," an upper limit on the number of established novelistic genres able to be supported by writers, readers, and publishers in any given year.

Finally, I explore the suggestion that generational changes in "mental climate" might manifest themselves not (only) in changes in novelistic genres, as suggested by Moretti, but rather in topical changes within novelistic genres. In studying a topic model (Steyvers et al., 2007) of the 7,800 volumes in UIUC's 19th century novels collection, I observe some evidence for topical trends cutting across multiple genres. For example, starting in the mid-19th century there is a proportional rise in a cluster of words suggestive of farming and rural life.

Moretti's work has been a touchstone for numerous discussions connected to the digital humanities—for example, the fate of "close reading". Perhaps more significant in the long-run is the interest his work has gained from scholars in the social and natural sciences, with examples ranging from formal reviews like that of statistician Cosma Shalizi (2006) to more informal commentary on quantitative literary history from social scientists like Henry Farrell (2010) and Andrew Gelman (2010). Finding scholars outside of literary studies and literary history publicly engaging with research is an important development. Given the desirability of new models of research in the humanities, Moretti's program can claim a interdisciplinary following that makes revisiting and extending his work of particular significance.

References

Bassett, Troy J. (2010). 'At the Circulating Library'. http://www.victorianresearch.org/atcl.

Farrell, Henry (2010). 'Hugo Awards II'. http://crookedtimber.org/2010/07/26/hugo-awards-ii/.

Garside,P. D., J. E.Belanger and S. A. Ragaz (2004). 'British Fiction,1800–1829: A Database of Production, Circulation & Reception'. http://www.british-#ction.cf.ac.uk/.

Gelman,Andrew (2010). 'The Triumph of the Thriller'. http://www.stat.columbia.edu/~cook/movabletype/archives/2010/03/the_triumph_of.html.

Moretti, Franco (2000). 'Conjectures on World Literature'. *New Lew Review.* .

Moretti, Franco (2003). 'Graphs, Maps, Trees, 1.'. *New Lew Review.* . http://newleftreview.org/?view=2482.

Moretti, Franco (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History.* London: Verso.

Shalizi, Cosma (2006). 'Graphs, Trees, Materialism, Fishing'. http://www.thevalve.org/go/valve/article/graphs_trees_materialism_fishing/.

Steyvers, Mark, Tom Griffiths, T Landauer, D Mcnamara, S Dennis , W Kintsch. (2007). 'Probabilistic Topic Models'. *In Handbook of Latent Semantic Analysis.* Lawrence Erlbaum Associates.

Notes

1. I owe the suggestion for a demographic approach to Shalizi (2006)'s review of *Graphs, Maps, Tress*: "There is a demography of businesses, of interest groups, even of medieval manuscripts of classical works, and so why not one of literary texts?"
2. The bibliographic databases used are Garside's *British Fiction Database, 1800-1829* with 2,272 titles and Bassett's *At the Circulating Library, 1837-1901* with 7335 titles
3. The methods presented are designed to accommodate changes to the underlying dataset and can be used by others who may disagree with classifications of particular novels and even with the bibliographic records provided by Garside and Bassett.

Computing in Canada: a History of the Incunabular Years

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta

Smith, Victoria Susan
victoriassmith@gmail.com
University of Alberta

Hoosein, Sophia
shoosein@ualberta.ca
University of Alberta

Gouglas, Sean
sean.gouglas@ualberta.ca
University of Alberta

Quamen, Harvey
hquamen@ualberta.ca
University of Alberta

1. Introduction

How were computers introduced to the public and how did humanities issues figure in the introduction of computing? The time has come in the digital humanities to think historically about computing in the humanities as Willard McCarty has pointed out in *Humanities Computing* and other venues. This paper describes a study of public representations of computing in Canada using the *Globe and Mail*, our major national newspaper. This paper restricts itself to what we call the incunabular years when computing was still a curiosity and business applications didn't yet dominate the public discourse.

References to digital computers in the Globe and Mail start in 1950 with a report of the annual meeting in London, England of Ferranti Ltd. This report describes under the heading of "Instruments" the development of a digital computer that was probably the predecessor of the Ferranti Atlas which pioneers like Susan Hockey worked on.

"The instrument department has a design team of considerable strength working in conjunction with Manchester University on the development of an electronic digital computer." (Globe and Mail, Oct. 31, 1950, p. 21)

From this first reference to digital computers buried in a business report, interest in and then anxieties about computers grow steadily through the 1950s and early 1960s until by 1964 the Globe runs a full page story in "The Woman's Globe and Mail" on "Will Computers Replace the Working Girl?" by Michelle Landsberg (Globe and Mail, May 21, 1964) that warns of the effects of automation on women who do most of the clerical work that can be automated. Computers go from being objects of curiosity in research labs at the University of Toronto with speculative utility to instruments that are changing the nature of office work, especially for women. It is this "incunabular" period that interests us, partly because it is a period when the academy is one of the major sites where computers are being installed and because academics are explaining computers to the broader community. This paper, using these early stories about computers, will tease out a history of early representations of computing in Canada. Specifically we will:

1. Describe the content analysis methodology used in this study.
2. Discuss the ways computers are presented to the public in the first decade and a half. Who is represented as having access to digital computers? What tasks are they presented as suitable for?
3. Discuss how computing jobs like data-entry, training and programming are gendered in the public discourse of the Globe.
4. Discuss the first references to the installation of computers at Universities and how these installations were presented to the public.
5. How was research computing presented and how were humanities applications of computing presented?

2. Methodology

The Globe and Mail online archive, unabashedly titled "Canada's Heritage from 1844," provides a full-text historical database of articles, ads, editorials and special features. In order to document the early discourse around computers we searched and collected all references to the word "computer" for content analysis coding and reading. The first reference we found dates back to 1897, though in this case it is not a reference to digital computers, but the computer as a type of job. The first reference to digital computers dates from 1950 and the number of references per year remains fairly low until the early 1960s when computing takes off as a subject of news, advertising and opinion.

Once gathered we coded the articles for content analysis. The coding rubric was developed iteratively as we read articles and developed hypotheses. For example, during the coding we became interested in gender and went back over the early stories to recode materials. Below is an example of the codes applied:

- Type of reference (i.e. news, classified ad, feature)
- Photos (i.e. were there photos and what do the photos show)
- Category of Application (University, Science, Military, Automation, Industry, Government, and Arts and Humanities)
- Gender of Named People (male, female, both, none)
- Discourse like "Brain" (Is the computer described as a giant "brain")
- Types of computers mentioned in the references where applicable

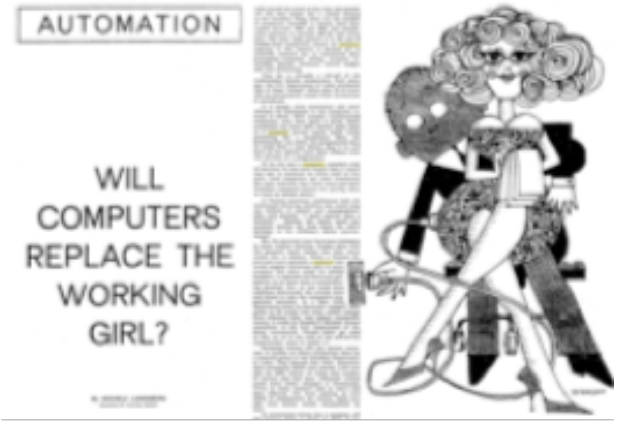
3. Themes Running Through the Early Years

Close reading and content analysis led us to identify and follow a number of themes running through the stories in these incunabular years. For example, related to the Ferranti mentioned in the first reference, is a Canadian turn from being oriented towards computing in the UK to being oriented towards computing from the USA. This turn is obviously a much larger issue for Canada after the war than just a change in where computers are coming from, but you see the turn in the articles from the 1950s. Bylines from London dominate in the early 50s, but by the end of the 1950s New York begins to be source of information about computers outside of Canada. You can see the early orientation towards the UK in titles like the 1955 article "Britain Leads in Office Automation." But, by 1961, in a comprehensive pair of articles on the computer industry, it is clear that US companies dominate. As the author Hugh Munro puts it, "Ferranti is the only company that designs and manufactures in Canada – specializing in big installations – which means the country is heavily dependent on the United States, where the other companies are based, not only for supplies but for technological advancement in the computer field." (Munro, "Surging Computer Industry Confident It Has Only Begun", December 6th, 1961, p. 23.)

4. Exclusionary Practices: Gender and Computing

One theme that stands out in the early representations of computing in the Globe is how women

were excluded from computing. Advertisements for programmers are in the “Help Wanted Male” section. It isn’t until 1960 that a woman is named in a story and then she is discussed as an exception. The exclusion of women gets discussed explicitly in 1964 when Michelle Landsberg writes the extraordinary feature “Will Computers Replace the Working Girl?” mentioned above (Globe and Mail, May 21, 1964). This feature confronts the effects of automation on women who happen to do most of the clerical work that is being automated. In the full paper we will contrast the exclusion with what we know of pioneers like Beatrice Worsley who was one two staff hired initially by the University of Toronto Computing Centre.



Computers and the Working Girl; Title, Illustration and Text

5. Computing in the University

The second story published about digital computers in the *Globe* is much more substantial than the Ferranti reference and it describes the first research computer installed in Canada at the University of Toronto. The story titled “Junior Electronic Brain Cost $100,000” dates from 1951, was accompanied by two photographs, and is about the UTEC Jr. computer installed at the U of T’s Computing Centre. It is really the first significant story about computing in the *Globe*, which is significant. In the full paper we will go into some detail about this first significant representation of computing as it illustrates many of the other points we want to make.

Much could be said about the U of T Computing Centre and the birth of academic computing in Canada. Here we will restrict ourselves to the way computing was presented to the public as starting at the University of Toronto. The U of T Computing Centre stayed in the news for decades, along with its director, Dr. Gotlieb, who was the most frequently mentioned expert of the period. But academic computing is not just about Gotlieb and research at his Centre. By 1955 we see the first ads for computing courses at U of T and the first reports of computers at other universities. By 1957 we see an article about computing in the arts and humanities. This article, titled, “Strange Music Made By an Electronic Brain” reports about a music composition experiment.

"To the casual observer, the squeaks, squawks, groans and hints of tunes were a harsh cacophony. To Professor C. C. Gotlieb and his colleagues, the sounds were the Iliac Suite, a string quartet composed by the electronic brain at the University of Illinois. It was an experiment in composition designed by Prof. Gottlieb and his fellow-workers with the university’s electronic brain to show that humans are not the only ones that can compose music."

What stands out about this and subsequent stories is that they are about the computer as an extraordinary device best understood as an “electronic brain” performing tasks that are human. The stories don’t really report humanities applications differently from scientific and engineering ones. Instead this research brain is presented as answering questions and completing tasks from all fields – it is a general purpose inquiry engine that the U of T Computing Centre is turning to questions and academic tasks from one field to the next. The brain is curious as we are, and the stories convey a sense of the discovery as Gotlieb’s team crosses disciplines discovering new uses for the computer. What remains to be seen (in a future study) is how the public discourse matches or reflects the evolving discourse within the academy and especially in humanities computing circles. We humanists, after all, are also reading the news; did we come to computing influenced by news of its promise or were we concerned about how it might affect our work?

Religo: A Relationship System

Rodríguez, Nuria
nro@uma.es
University of Málaga (Spain)

Isolani, Alida
isolani@signum.sns.it
Scuola Normale Superiore (Italy)

Lombardini, Dianella
dianella@signum.sns.it
Scuola Normale Superiore (Italy)

Marotta, Daniele
marotta@signum.sns.it
Scuola Normale Superiore (Italy)

1. Introduction

Over the years, digital libraries and textual archives have collected, described, and classified texts and multimedia objects. These kinds of repositories are effective in compiling, describing, and disseminating the cultural heritage such as the artistic and literary expressions. Also, many of them, following the developments of Computational Linguistics, have incorporated tools for textual analysis as part of their end-user services. Nevertheless, these systems are weak in terms of *relationships*. Of course, they are configured in such way that it is possible to relate the digital objects compiled; thus, for instance, it is easy to retrieve a set of visual artefacts sharing the same subject matter. However, these relationships are based on the traditional criteria of classification and description (metadata and keywords), without any intention of exploring the nature or specific characteristics of the relationships that the Art History discipline's phenomena maintain among them.

We should not overlook the intrinsic relationship that exists among texts, concepts – or ideas –, words and visual artefacts in the construction of art-historical knowledge (Mitchell, 1994). As Heffernan (2006) argues for the case of words and images, this relationship should not taken as a simple reproduction of art works by a set of words, but rather the conjunction of visual artefacts and words generates new knowledge. This is one of the reasons for which we can contend that these relationships deserve to be treated as a research object in themselves.

Therefore, our intention is to explore the potential relationships that could be established among these entities: texts, concepts, words, and visual artefacts with the aim of investigating how these relationships are able to produce new significant knowledge or are able to open new understandings.

2. Religo

Religo is a system that enables the construction of interpretations based on relationships. According to the researcher's needs and the art historian's research habits and procedures, the question of *text and image* induces us to design a system to establish relations between various kinds of objects (texts, images, videos, etc.), to provide new possibilities for analysis and research, which are offered only partly by the state of the art.

Currently, in the most significant projects[1] working on texts and images – although with interesting and useful features – some limitations can be found in terms of:

- functionalities: only tagging or object manipulation;
- usability: complex and without a user-friendly interface;
- purpose: oriented to social participation (this feature often does not reach the entire scholarly community).

Taking into account this state of the art, Religo does not reduce the relationship to the concept of pure tagging or to the idea of simple connection between entities, but treats it as basic element for interpretation and analysis, making itself the subject of research in order to create new knowledge.

Religo relates the domain entities creating two logical levels: the expression one, consisting of *digital objects* on which the interpretation can be developed; and the semantic one, consisting of *digital concepts* (the relationships between digital objects and the predications on themselves) that allow the interpretation to be built (Buzzetti, 2004)[2].

This means that, when a digital concept is the subject of interpretation, it is placed on the expression level, becoming a digital object itself.

An example is the Michelangelo's masterpiece *The Last Judgement*, analysed and interpreted by the Spanish author Francisco Pacheco in his 17th treatise *The Art of Painting* (1649). In this case, *The Last Judgement* would be the digital object, and the interpretation given by Pacheco, the digital concept. However, insofar as the Pacheco's interpretation is

also subject of interpretation and analysis by the modern historiography, it in turn becomes a digital object.

The entities of the domain can be submitted to a number of general operations (such as selection of parts, links, free tagging or metadata encoding, etc.) and others more specific according to their particular features (specific operations on texts and image).

In every operation the centrality of the relationship is clear: from its creation, made easily by a simple *drag and drop* of selected portions of objects, up to reach the composition of documents as result of the different entities relationship that themselves constitute the new knowledge of the study process.

A more evident utility of this new use of relationships can be seen in the search and navigation functions, for example to improve the search capability because it ensures a higher degree of *precision* and *recall*[3].

As an example, searching for the word *emblema*, Religo returns both entities containing *emblema* as textual occurrence (W) and those where *emblema* does not appear (NW) but which are related to W (Figure 1). Thus, the relationship gives relevance and importance to the entity NW, which otherwise, from a purely textual standpoint, would go unnoticed or simply would not exist.

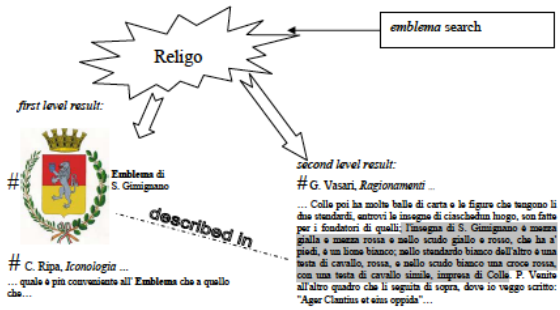


Figure 1[4]

Relationship also affects the display of the entities: that is, in addition to the classical view *as list*, Religo provides a view *as graph*, creating a network between the various domain entities which on one hand allows the reconstruction of interpretive reading by simply moving the focus between digital objects, on the other hand enables a contextualized vision of every digital object. These features, typically used during the work process, are also useful in order to share and exploit the research results.

As a more concrete example, let us consider the following domain:

The Art of Painting by Pacheco discusses Michelangelo and his works, including *The Last Judgement*. It also treats other painters such as Velázquez or Tiziano, other works such as *The Final Judgement* painted by Pacheco himself, the portrait of King Phillip II painted byTiziano, and other artists such as Dolce, Paleotti, Lomazzo, or Céspedes.

Religo allows relationships to be created between these objects:

[*Pacheco*] **author of** [*The Art of Painting*]

[*Michelangelo*] **painter of** [*The Last Judgment*]

[*The Last Judgement by Michelangelo*], [*The Last Judgement by Pacheco*] **described in** [*The Art of Painting*]

[*Phillip II by Tiziano*] **mentioned in** [*The Art of Painting*][5]

[*The Last Judgement of Michelangelo*] **influences on** [*The Last Judgement of Pacheco*]

[*The Last Judgement of Michelangelo*] **used as example by** [*Dolce, Pacheco, Lomazzo, Céspedes*]

[*Dolce, Paleotti, Lomazzo, Céspedes*] **cited in** [*The Art of Painting*]

[*The Last Judgement of Michelangelo*] **illustrates concepts of** [*deviations of decoro, terribilità, movements and affetti*]

[*The Last Judgement of Michelangelo*] **described with terms** [*artificioso, espantoso, terrible, horribilidad, feroz*]

These relationships themselves become new entities of the domain, forming an interconnected network and producing the following new knowledge level:

[*Lomazzo*] [*Céspedes*] **cited by** [*Pacheco*] **to define the concept of** [*painting*]

At this level, we can see how the result specifies the connection among the different theorists cited by Pacheco and the concepts that he defines in *The Art of Painting* until creating another richer level:

[*The Last Judgement*] **used as example by** [*Pacheco, Dolce*] **to illustrate idea of** [*deviations of decoro*]

[*The Last Judgement of Michelangelo*] **used as example by** [*Pacheco, Lomazzo*] **to illustrate the idea of** [*movements and affetti*]

[*The Last Judgement of Michelangelo*] **used as example by** [*Pacheco, Céspedes*] **to illustrate the idea of** [*terribilità*]

What we can deduce from this result is that, from the second half of 16th century, *The Last Judgement* by Michelangelo plays the role of universal reference to illustrate or exemplify a wide range of aspects concerning the visual arts, being used by each author in a different way. *The Art of Painting*, as an encyclopaedic treatise, brings together many of these interpretations, which Pacheco unifies into a single point of view.

The most interesting results arise when we use a complete repository of works, images, and texts. For example, if we consider a repository of Spanish 17th treatises, as ATENEA Project[6], we might find the following types of relationships:

[*The Last Judgement of Michelangelo*] **described by** [*Pacheco*, 1649] and [*Carducho*, 1634] only **mentioned by** [*Martínez*, ca. 1675]

[*The Last Judgement of Michelangelo*] **described by** [*Pacheco*] and [*Carducho*] **with coinciding terms** [*confusión, temor, horribilidad, terrible*].

3. Conclusions and Future Developments

As an initial task, Religo is provided with all the typical features to operate on texts and images in terms of combination of interacting tools for example to describe and catalogue visual artefacts, to analyse images, to manipulate images, or to annotate images (whole or partly).

Together with the standalone version, an online should be allowed in order to ensure content sharing and social tagging in expert contexts of usage.

Moreover, the system would be generalized for use in other different domains and would have the capacity to handle other types of entities such as audio and video.

References

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

Heffernan, J. (2006). *Cultivating picturacy: visual art and verbal interventions.* Waco, TX: Baylor University Press.

Mitchell, W.J.T. (1994). *Picture Theory: Essays on Verbal and Visual Representation.* Chicago: Chicago University Press.

Notes

1. Image Markup Tool `http://tapor.uvic.ca/~mholmes/image_markup/`. Pinakes `http://pinakes.imss.fi.it/index.php/Main_Page`. VLMA `http://lkws1.rdg.ac.uk/vlma/`. EPPT `http://beowulf.engl.uky.edu/~eft/eppt-trial/EPPT-TrialProjects.htm`. Flickr `http://www.flickr.com/`. TextGrid `http://www.textgrid.de/`. Talia `http://net7sviluppo.com/trac/talia/wiki/TaliaSystemDescription`.
2. This theoretical model has been used by Signum to develop a system for facilitating semantic research and text reading in Text and Semantics `http://textandsemantic.signum.sns.it`.
3. *Precision* can be seen as a measure of exactness or fidelity, whereas *recall* is a measure of completeness.
4. C. Ripa, *Iconologia, overo Descrittione dell'Imagini universali* […], Roma 1593, p. 96. G. Vasari, *Ragionamenti di Giorgio Vasari pittore ed architetto aretino* […], Firenze 1832-38, p. 1404.
5. Notice the difference between *describe*, which implies a detailed explanation of the painting; and *mention*, which only means that the painting has been cited.
6. `http://www.proyectoatenea.es`

Development of Digital Projects as Learning Strategies. The Desingcrea/Diseñoteca Project

Rodríguez, Nuria
nro@uma.es
University of Malaga (Spain)

In the last Digital Humanities conference in London (July 2010), several of those present expressed the need for educational strategies based on digital projects. The reasons can be summarised as follows: firstly, given the difficult sustainability of these types of projects, the help of students in their development could be a key factor in their upholding. Secondly, participating in real digital projects could help these students gain an array of essential competences in the digital society in which they will develop their professional work.

Bearing this in mind, the aim of this paper is to present a case of educational innovation, financed by the University of Malaga since 2006, which I believe can respond to the concerns expressed in London.

2. The Desingcrea/Diseñoteca Project

The Desingcrea/Diseñoteca project was originally created to redefine the practical aspect of certain subjects within the degree of Technical Industrial Design Engineering given in the Polytechnic Institute of the University of Malaga.

Specifically, our aim was to develop collective learning strategies that could lead to significant and relevant training, involving the student in the development of his/her own knowledge. In order to do this we decided that the new possibilities of virtual interaction brought about by the Information and Communication Technologies, as proposed in the Technology Enhanced Learning (TEL) theory, social participation web environments (2.0), combined with the educational principles of Pedagogic Constructivism (Mitchel Rescnick, 1996) and Conectivism (G. Siemens, 2006), offered the perfect framework on which to base this project.

Taking an objective based educational model as a starting point, we decided that all these strategies revolved around a common objective: the development of an industrial design database, which we called Diseñoteca, which in turn would be integrated into a web portal also dedicated to industrial design (Desingcrea). The idea was that Desingcrea/Diseñoteca would be developed progressively through the collective work of students, who would be responsible for the preparation of its material and content, through tasks incorporated into the teaching programme of certain subjects of the degree. Once developed, the Desingcrea/Diseñoteca system would work as an online resource available to all students, who could use it as a tool for consultation, study and debate.

3. Desingcrea/Diseñoteca: Design and Structure

Desingcrea/Diseñotca is a web application developed specifically by computer engineers from the University of Malaga for this project1. Updates carried out on the system aimed to transform an informational system, based on the «database» philosophy, into a learning community, based on the «social network» or «2.0» theory, bringing together users interested in industrial design, amongst them our students, who were responsible for its dynamization.

Desingcrea/Diseñoteca focuses on the industrial design, therefore the entire information published on this website is around this subject matter. As well as informative sections, created by the students, such as directories, bibliographical indexes, etc., the Desingcrea portal has resources common to web 2.0, such as the publication of news and articles as blogs, the possibility for users to evaluate the contents, RSS, tag clouds and the capacity to link these contents to other existing resources on the Network (text, images, web pages, etc.). It is therefore possible to create a real, open and global network about industrial design, which students use as a personalised learning environment. [Fig. 1 and 2].

Fig. 1. Desingcrea website. 1 <designcrea.uma.es>

Fig. 2. Blogs written by students at Desingcrea website.

The Diseñoteca database, which is integrated into the Desingcrea website, is also being developed through the collective work of students. Industrial design objects are registered in a structured manner within Diseñoteca. It is made up of data records in which participants describe and classify the design objects according to a protocol of standards and metadata. [Fig. 3]

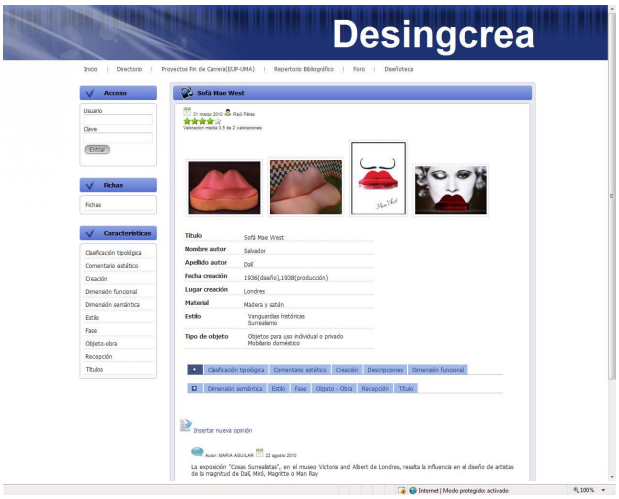


Fig. 3. Diseñoteca' s record. Simple display showing the assessment given by students.

In order to create this data register, the principal source used was The Categories for the Description of works of Art (CDWA), of the Getty Research Institute (1998), which was updated in the work Cataloging Cultural Objects (Baca, 2006), of American Library Association Editions. The result is a more specific example of this standard that we call Categories for the description of industrial design objects (CDOID), and that will be the object of a future publication.

The system offers students the possibility to discuss the records of their classmates, contributing data and information that enrich the description of the object. Besides, in this way the collective participation in the creation of contents and the exchange of ideas is increased.

Likewise, students can assess these records using tools common to the repositories of the social web. This assessment will act as a co-evaluation, complementing the final grade given by the teacher. [Fig. 3].

Management and monitoring of students. Each teacher is assigned a specific number of students to monitor. A series of features have been installed on the system so that students are monitored in the most complete way possible. In this way it is possible for the teacher to see the records that have been created and edited by any student and consult his/her record history, seeing the different actions and tasks that have been carried out. This way the system not only shows us the amount of time that the student has been connected to the Internet, but also the tasks that he/she has actually carried out and therefore if the student has really been working. [Fig. 4].

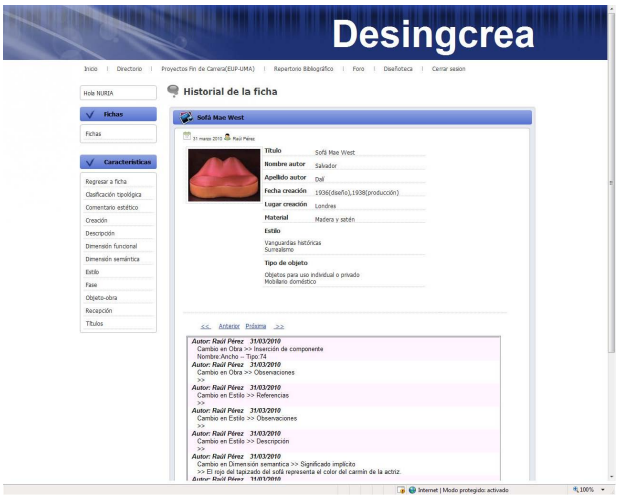


Fig. 4. Display for the teacher profile with the report of the student´s work and actions.

Finally, given that one of the aims of Desingcrea/ Diseñoteca, as mentioned, is to work as an information and meeting website for all those interested in industrial design, a validation tool has been incorporated into the system as a means to guarantee the quality of the process. That is, information introduced into Desingcrea/Diseñoteca is not made public to users until it has been corrected and validated by the teacher in charge.

4. What does Desingcrea/Diseñoteca Bring to the Learning of Students?

According to results obtained since this project began in 2006, I list the following aspects which could be applied to any similar, digital based projects.

- The students have become familiar with work concepts and systems common to the design and execution of digital projects, offering them an invaluable preparation in facing this type of project in the future. For example, they have become familiar with the use of metadata, and have become aware of the need for their systematic and coherent use, an idea that is not always sufficiently adopted amongst humanities specialists.
- With regards to the pedagogic use of 2.0 tools, the potential of blog type applications to encourage critical thinking should be underlined. News and articles published by students, as contents on the Desingcrea portal, can be discussed and assessed by their own classmates, favouring the exchange of ideas and encouraging debate amongst them.
- The possibility of linking information published on Desingcrea/Diseñoteca with other news, portals and sites on the Network, helps students to familiarise themselves with the idea of the Network as a global data repository where knowledge is produced through the significant connection of information nodes.
- While creating Designcrea/Diseñoteca, a collaborative project based on the contributions of participants on an open system which, at the same time, is incorporated globally into the Network from which it is also fuelled, students are encouraged to integrate themselves into the so called open and shared knowledge culture. This culture is associated to the thinking behind 2.0 and defines the new paradigm of knowledge that, based on the concepts of both Rheingold's Smart Mobs (2000) and Lévy's Collective intelligence (1994, 1998), is shaping our contemporary society (UNESCO, 2005).

During the paper, these ideas will be developed in more detail and other significant results will be presented.

References

Baca M. (ed.) (2006). *Cataloging Cultural Objects: A Guide to Describing Works and Their Images.* Chicago: American Library Association Editions.

Cabero Almenara, J. (2007). *Diseño y producción de TIC para la formación: nuevas tecnologías de la información y la comunicación*. Barcelona: UOC.

Fisher, D. L., KHINE, M. S. (2006). *Contemporary approaches to research on learning environments: worldviews.* New Jersey: World Scientific.

Lévy, P. (1994). *L'intelligence collective. Pour une anthropologie du cyberespace.* París: La Découverte.

Lévy, P. (1998). *L'intelligence collective, une nouvelle utopie de la communication?* Available at: `http://membres.lycos.fr/natvidal/levy.htm`.

Litwin, E. (2005). *Tecnologías educativas en tiempos de Internet.* Buenos Aires, Madrid: Amorrortu.

Mitchel Resnick, Y. K. (1996). *Constructionism in practice: designing, thinking, and learning in a digital world.* Mahwah (New Jersey): LEA.

Rheingold, H. (2002). *Smart Mobs. The Next Social Revolution.* Cambridge: Perseus Publishing.

Seely Brown, J. (2002). 'Growing Up Digital: How the Web Changes Work, Education, and the Ways People Learn'. *Change.* 10-20.

Siemans, G. (2006). *Knowing Knowledge* Available at `http://www.knowingknowledge.com`.

UNESCO (2005). *Hacia las sociedades del conocimiento*, Ediciones UNESCO.

An Ontological View of Canonical Citations

Romanello, Matteo
matteo.romanello@kcl.ac.uk
Kings College London

Pasin, Michele
michele.pasin@kcl.ac.uk
Kings College London

Canonical citations are references to Classical (i.e. Greek and Latin) texts that are expressed by scholars by means of an abridged canonical format (Romanello 2008; Romanello 2007). They fulfil the function of providing an abstract reference scheme for texts (somehow similar to the function of geographical coordinates to express references to places) since they allow us to express references to them no matter what particular text edition we are actually looking at. For example, the reference to the twelfth book of Homer's *Iliad* expressed as “Hom. *Il*. XII 1” can be resolved to the text of the same passage as established in various critical editions of that work, where “XII” and “1” are the “coordinates” that allow us to locate that precise text passage within all critical editions of Homer's *Iliad*.

Historically, canonical references are the result of an effort – whose origins can be traced back to the Renaissance (Martin 2003; Berra 2011) – made by the scholarly community as a whole to provide a precise, stable and shared way to refer to Classical texts. Since the early stages of Humanities Computing and Digital Humanities (Bolter 1993; Crane 1987; McCarty 2005), canonical references were regarded as the ideal candidate on which to experiment the potentialities of hypertext: indeed they can be seen as hyperlinks *in potentia* pointing a text from within another. More recently (Crane et al. 2009) they were considered as a discipline-specific kind of named entities that Classics scholars should be provided with tools to search for within their texts.

J.D. Bolter describes classical philology as “the art of explicating an ancient text by exploring its relationships to other specific texts and to the corpus of ancient literature as a whole”. In such a discipline the act of *referring* to texts – that J. Unsworth has listed among the “scholarly primitives” (Unsworth 2000) – becomes even more crucial than in other disciplines since texts are the very research objects of classical philology and references to them play a key role in constructing argumentations.As N. Smith (2009) has already pointed out, canonical citations reflect an ontological view of texts in this specific domain and specifically how classicists perceive ancient texts as objects. In this paper we present the Humanities Citation Ontology (HuCit)[1], an ontology that aims at characterising the semantics of citations as they are normally conceived and used in humanistic disciplines. We claim that the specification of such an ontology is worthwhile for at least the following reasons:

1. it allows us to disentangle from an ontological point of view the complex relationships between, for instance, a canonical reference found in a journal paper and the manuscripts and editions of the text that we can access via that reference;
2. it allows us to define types of references and alternative representations of the same reference: this is an important step towards tools that allow automatic formatting of such references according to various styles (as it happens already for modern bibliographic references with Zotero);
3. it provides us with a way to access the meaning of canonical references beyond their surface appearance, which might vary substantially as in the case of “Hom. *Od*. I 1” and “α 1”, two canonical references to the same passage but conforming to different citation styles.From an initial analysis of the relevant literature we concluded that none of the existing ontologies actually model the deep meaning of canonical references. An interesting attempt to formalise citations by means of an ontology is CITO (Shotton 2010) which however looks exclusively at modern bibliographic references and focuses in particular on citation types. As it was observed already (Smith 2009; Mimno et al. 2005), the distinction made by FRBR (Functional Requirements for Bibliographic Records) between a work, its expressions and its manifestations can be adapted to represent texts in the Classics domain as well. In this paper we propose an initial implementation of a canonical reference ontology based on FRBRoo which is the result of a process aimed at harmonising FRBR with the CIDOC Conceptual Reference Model (CIDOC-CRM) (Doerr & LeBoeuf 2007).

A key aspect we have to face is to determine at which ontological level of the cited object a canonical reference is pointing to. A citation such as “Hom. *Od*. I 1” is it referring to the abstract notion of *Odyssey* (i.e. a *work* in the FRBR model) or to a particular version (e.g. edition, translation, etc.) of that work (i.e. a FRBR expression)? It might help to

observe that this reference can be solved by a human reader for example into the text of that passage in French translation: therefore it is not being specified at the expression level. The textual coordinates of the citation, namely “first line of the first book”, expressed by the string “I 1” clearly refer to a logical citation scheme that applies already to the abstract notion of *Odyssey* (i.e. a FRBR work). Thus we can say that a canonical citation follows a given citation scheme that characterises a particular literary text and might differ from one to another. That citation scheme is a conceptual object and is the result of the work of scholars to guarantee the ability of citing literary texts.

To illustrate the notion of *logical* citation scheme as opposed to a *physical* one let us examine a single case, that is the Athenaeus' *Deipnosophistae*. Scholars cite this work by means of canonical references that follow a logical citation scheme derived from a physical one (e.g. “Ath. *Deipn*. XV 694 e-f”). The textual coordinates “694 e-f” refer to the pagination of the edition of the text by Isaac Casaubon dated 1598, and specifically to sections “e” to “f” of page 694 of that edition. At first it seems a physical citation scheme. But since all editions after Casaubon's provide the readers with marginal numbers referring to that pagination it became a logical citation scheme: indeed 694 does not refer anymore to a physical page within more recent editions such as Olson's.

Canonical citations have both form and content. Different citations might differ by form but they can still have the same content. The content of a citation is the abstract reference of which that citation is an expression. For example, a citation to the first line of Homer's *Iliad* can be written in several ways according to different citation styles. Nevertheless “Hom. *Il*. I 1”, “A 1” and “Homer, *Iliad* 1.1” are different expressions of the same canonical reference to Homer's *Iliad* (cited by book and then by line). Given all these reasons, we propose to introduce - among the others - the classes Citation, Reference and Citation_Style to the "E28 Conceptual Object" branch of CIDOC-CRM (we will discuss the details of this approach at the conference, also in the light of most recent activities to harmonise CIDOC and FRBR). Further work is then required in order to extend this conceptual model so that it can support more complex reasoning tasks, such as translation mechanisms among different citation schemes, or the automated extraction of citations from non structured materials.

To sum up, in this paper we describe the implementation of an ontology to model canonical references that builds upon the solid conceptual models already defined by CIDOC-CRM and FRBRoo. In the framework of a Classics cyberinfrastructure (Crane et al. 2009), such an ontology is meant to support the interoperability of tools that are being currently developed to extract (Romanello et al. 2009), retrieve (Smith 2009) and solve (Ruddy & Rebillard 2009) canonical references.

References

Berra, A. (2011). 'Manier le thésaurus grec.'. *Les main de l'intellect. Lieux de savoir..* C. Jacob (ed.). Paris: Albin Michel.

Bolter, J.D. (1993). 'Hypertext and the Classical Commentary.'. *Accessing antiquity : the computerization of classical studies.* Tucson: University of Arizona Press, pp. 157-171.

Crane, G. (1987). 'From the old to the new: intergrating hypertext into traditional scholarship.'. *Proceedings of the ACM conference on Hypertext.* Chapel Hill, North Carolina, United States, pp. 51-55. http://doi.acm.org/10.1145/317426.317432 (accessed February 2, 2009).

Crane, G., Seales, B. & Terras, M. (2009). 'Cyberinfrastructure for Classical Philology.'. *Digital Humanities Quarterly.* 3(1). http://www.digitalhumanities.org/dhq/vol/3/1/000023/000023.html (accessed July 19, 2010).

Doerr, M. & LeBoeuf, P. (2007). 'Modelling Intellectual Processes: The FRBR - CRM Harmonization.'. *Digital Libraries: Research and Development. Lecture Notes in Computer Science.* C. Thanos, F. Borri, & L. Candela, eds. (ed.). Berlin / Heidelberg: D. Springer, pp. 114-123. http://dx.doi.org/10.1007/978-3-540-77088-6_11.

Martin, H. (2003). 'Du livre a la lecture.'. *Des Alexandries II. Les métamorphoses du lecteur.* C. Jacob (ed.). Bibliothèque nationale de France, pp. 35-45. http://hal.archivesouvertes.fr/hal-00131623/en/ (accessed October 26, 2010).

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

Mimno, D., Crane, G. & Jones, A. (2005). 'Hierarchical Catalog Records'. *D-Lib Magazine.* 11(10). http://www.dlib.org/dlib/october05/crane/10crane.html (accessed July 22, 2010).

Romanello, M. (2008). 'A semantic linking framework to provide critical value-added services for Ejournals on classics.'. *ELPUB2008. Open Scholarship:*

Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing. S. Mornati & L. Chan, eds. (ed.). Toronto, Canada, 25-27 June 2008, pp. 401-414. http://elpub.scix.net/cgi-bin/works/Show?401_elpub2008 (accessed August 11, 2008).

Romanello, M. (2007). 'A Semantic Linking System for Canonical References to Electronic Corpora.'. *International Conference on Electronic Corpora of Ancient Languages : proceedings of the international conference.* P. Zemanek (ed.). Prague, November 16-17, 2007, pp. 107–120. http://eprints.rclis.org/16239/1/Romanello2008.pdf.

Romanello, M., Boschetti, F. & Crane, G. (2009). 'Citations in the Digital Library of Classics: Extracting Canonical References by Using Conditional Random Fields.'. *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries.* Suntec City, Singapore: Association for Computational Linguistics, pp. 80–87. http://portal.acm.org/ft_gateway.cfm?id=1699763&type=pdf.

Ruddy, D. & Rebillard, E. (2009). 'Text Linking in the Humanities: Citing Canonical Works Using OpenURL.'. . http://www.cni.org/tfms/2009a.spring/abstracts/PB-text-ruddy.html (accessed September 11, 2009).

Shotton, D. (2010). 'CiTO, the Citation Typing Ontology.'. *Journal of Biomedical Semantics.* 1(Suppl 1): S6. http://dx.doi.org/10.1186/2041-1480-1-S1-S6 (accessed October 25, 2010).

Smith, N. (2009). 'Citation in Classical Studies.'. *Digital Humanities Quarterly.* 3(1). http://www.digitalhumanities.org/dhq/vol/003/1/000028.html# (accessed March 15, 2009).

Unsworth, J. (2000). 'Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?'. http://www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html.

Notes

1. http://purl.org/net/hucit

Alma Cardell Curtin and Jeremiah Curtin: the Translator's Wife's Stylistic Fingerprint

Rybicki, Jan
jkrybicki@gmail.com
Pedagogical University of Kraków, Poland

1. The Problem

Poland's first literary Nobel Prize winner, Henryk Sienkiewicz (1846-1916), owed his great if short-lived fame to the very numerous if very mediocre translations by Jeremiah Curtin (1835-1906), diplomat, ethnographer, polyglot. Born in a Catholic Irish family in Wisconsin, he graduated from Harvard and was posted as a secretary to the American mission in St. Petersburg, Russia; his fluency in Russian made him a popular figure among the local aristocracy. Paradoxically, it might have contributed to his conflict with Ambassador Clay and precipitated the end of his diplomatic career (1869). Curtin switched to two professions he continued till the end of his life: that of the ethnographer (employed, for a time, by the Smithsonian Institution) and of the literary translator. In 1872, he met and promptly married Alma Cardell (1847-1938), who soon abandoned her post as a teacher in a Soldiers' Orphans' Home (for which her studies at the Barre Academy had made her more than qualified) to become her husband's secretary, amanuensis and editor; until her peripatetic husband's death, her life was to be led in hotels and boarding houses around the globe – especially after the Curtins stroke gold with Sienkiewicz's international bestseller *Quo vadis* (1896). Alma devoted much work to virtually all publications signed by Jeremiah: his translations of Sienkiewicz and of other Polish authors (Orzeszkowa, Prus and Potocki); his translations from the Russian (Gogol, Zagoskin's and Alexy K. Tolstoy); and his ethnographic studies on myths of Native Americans, Ireland and Slavic peoples. She also published and edited three books on Mongols after her husband's death.

Yet the story of Sienkiewicz's translator is most extensively told in *Memoirs of Jeremiah Curtin* (1940), published after the death of Alma Cardell Curtin. Although written in first-person narrative, they have been since proven to be the work of the wife. Michael Jacek Miko? has shown that Curtin's alleged

memoirs are in fact a compilation of “somewhat edited" fragments of Alma’s diaries and letters to her family (Miko? 1990). The same diaries (and not the *Memoirs*) show the extent of Alma’s contribution to Jeremiah’s translations: after a whole day of taking her husband’s dictation, he would go to sleep while she would copy and correct the day’s work (Miko? 1994). And although she knew no Polish and, at most, but a little Russian, it is not implausible to suspect that some traces of the translator’s wife’s hand might have been left on Sienkiewicz’s fiction in English. In the most radical hypothesis by Cheryl L. Collins, Alma, “held hostage by Jeremiah’s almost pathological restlessness," could have been “his full partner" in his literary work (Collins 2008).

This presents a nice authorship attribution problem. All that could have been done in this respect with traditional methods has been done by Miko?; the rest is the attributor’s nightmare, as all work published under Curtin’s name has been preserved in manuscripts in Alma’s hand alone. And while traces of the style of the *Memoirs* could perhaps be found in Alma’s editions of Jeremiah’s ethnographic works, it is highly uncertain if similar traces can at all be found in his translations.

2. The Method

All hope there is lies in non-traditional methods of authorship attribution, developed at least since the seminal *Inference and Disputed Authorship: the Federalist* (Mosteller, Wallace 1964) and proven to be helpful in plagiarism detection (although Alma’s is plagiarism *a rebours*). This study applies Cluster Analysis to normalized word frequencies in texts; as shown by (to name but a few) Burrows (1987, 2002), Hoover (2004, 2004a) or Daren-Oskam (2007), this is one of the most precise methods of “stylistic dactyloscopy." A script by Maciej Eder, written for the R statistical environment, converts the electronic texts to produce complete most-frequent-word (MFW) frequency lists, calculates their z-scores in each text according to the Burrows Delta procedure (Burrows 2002); selects words for analysis from various frequency ranges; performs additional procedures for better accuracy (Hoover’s culling and pronoun deletion); compares the results for individual texts; produces Cluster Analysis tree diagrams that show the distances between the texts; and, finally, combines the tree diagrams made for various parameters (number of words, degree of culling) in a bootstrap consensus tree (Dunn et al. 2005, quoted in Baayen 2008: 143-147).

The analysis included all original works by Curtin (12 extensive studies) and a great majority of his translations (21 novels or long novellas).

Figure 1. shows the patterns of similarity and difference of word frequencies in all texts studied. The bootstrap consensus tree neatly divides Curtin’s *oeuvre* into two discrete groups: the upper branches are his translations while his *Memoirs* and his ethnography lie below. Yet the *Memoirs* are placed away from his ethnographic studies; also, what has been proven by Miko? to be Alma’s work, lies close to two of Jeremiah’s books on the Mongols, published by Alma after his death.

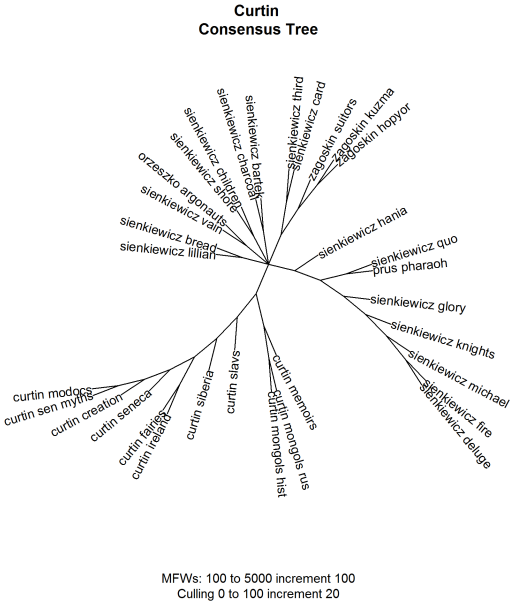


Figure 1. Consensus tree for all texts

The length of the word list used in this study (all the way to the 5000th most frequent word in the corpus) can raise doubts as to the validity, in this context, of the term *stylistic* similarity: after all, words so far down the frequency-ordered word list are often quite meaningful and might reflect differences of content as well as of style. Any chance of finding traces of the translator’s wife’s hand in the translations as well as in the *Memoirs* has to rely on somewhat shorter lists (tentatively, from 10 to 150) from the top of the frequency-ordered most-frequent-word list; dominated as it is by functions words and aided by a 100% culling rate (which limits the analysis to words that appear in all texts studied) and personal pronoun deletion, it might help purge any impact of the texts’ content to try to bring, say, a book on Irish myths and a translation of Sienkiewicz’s historical romance, to a common stylistic, or perhaps simply lexical, denominator.

The new result is presented in Figure 2. The texts attributed to Mrs Curtin either by Miko? or by Figure 1. are still immediate neighbours; more interestingly, the *Memoirs* (and the two "Mongol" works) now grow on the same branch with two translations: *Lillian Morris* and *For Bread*.

Figure 2. Consensus tree for short most-frequent-word lists and a 100% culling

3. Conclusions

So far, attempts at finding stylistic traces of the translator or the editor have been only partially successful. Word-frequency-based stylometric methods have shown that they are better at attributing the author of the original than the translator (Rybicki 2009, 2010; Rybicki, Eder 2010), or that Henry James's switch from writing to dictation left no stylistic trace (Hoover 2010).

This time, however, it is possible that the translator's wife has left her mark. First, Miko?'s attribution of the *Memoirs of Jeremiah Curtin* is visible in the consensus tree diagrams; second, the diagrams seem to validate the suspicions of Mrs Curtin's significant contribution to *The Mongols: A History* and *The Mongols in Russia*.

Even more importantly, the most frequent words in two of the Sienkiewicz translations seem to exhibit a greater affinity to the *Memoirs*. It is hardly a coincidence that the two short stories depart from Sienkiewicz's usual male- and Poland-dominated writing. Indeed, *Lillian Morris* and *For Bread* are two of Sienkiewicz's so-called "American stories" (written during or immediately after his stay in America, 1876-1879) and both feature a strong female protagonist. If it ever made sense for Curtin to give his wife a somewhat freer hand at shaping the final version of the English version of this or that Sienkiewicz text, it would have been on the occasion of these two stories. After all, during the work on *Lillian Morris* and *For Bread* in the Irish summer of 1893, Jeremiah was busy, struggling with several other projects at the same time: he was collecting material for one of his Irish mythologies, translating a couple of other short works Sienkiewicz, and writing a lengthy "Translator's Preface' to the Polish writer's another historical romance, *Pan Michael*.

References

Baayen, R.H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics using R.* Cambridge.

Burrows, J.F. (1987). *Computation into Criticism: a Study of Jane Austen's Novels and an Experiment in Method.* Oxford.

Burrows, J.F. (2002). 'Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship'. *Literary and Linguistic Computing.* 17: 267-287.

Collins, C.L. (2008). 'Behind the Curtin'. *Milwaukee Magazine.* 1. April.

Curtin, J. (1940). *Memoirs of Jeremiah Curtin.* Madison.

Dalen-Oskam, K. van, Zundert, J. van (2007). 'Delta for Middle Dutch—Author and Copyist Distinction in Walewein'. *Literary and Linguistic Computing.* 22: 345-362.

Dunn, M. , Terrill, A. , Reesink, G. , Foley, R.A., Levinson, S.C. (2005). 'Structural Phylogenetics and the Reconstruction of Ancient Language History'. *Science.* 309: 2072-2075.

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. (2010). 'Modes of Composition in Henry James: Dictation, Style, and What Maisie Knew'. *Proceedings of the Digital Humanities 2009 conference.* College Park, MD.

Miko, M.J. (1990). 'Alma Cardell Curtin, the Woman behind Jeremiah Curtin'. *Milwaukee History.* 13: 53-68.

Miko, M.J. (1994). *W pogoni za Sienkiewiczem.* Warszawa.

Mosteller, F., Wallace, D.L. (1964). *Inference and Disputed Authorship: The Federalist.* Reading.

Rybicki, J. (2009). 'Liczenie krasnoludków. Troch? inaczej o polskich przek?adach trylogii Tolkiena'. *Ludzie i krasnoludki – powinowactwa z wyboru? Conference proceedings.* Warszawa.

Rybicki, J. (2009). 'Translation and Delta Revisited: When We Read Translations, Is It the Author or the Translator that We Really Read?'. *Proceedings of the Digital Humanities 2009 conference.* College Park, MD.

Rybicki, J., Eder, M. (2011). 'Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?'. *Literary and Linguistic Computing.* 26: (forthcoming).

Evaluating Digital Scholarship: A Case Study in the Field of Literature

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

Mandell, Laura
laura.mandell@gmail.com

Olsen, Stephen
SOlsen@mla.org

Evaluating and receiving credit for digital scholarship within traditional disciplinary areas in the humanities has been a concern much discussed, not only within the digital humanities community, but at think tanks on the future of scholarly publishing, within institutions, at professional associations in the various disciplines of the humanities, and in journal and newspaper articles.[1]

Over the past few years The Modern Language Association has taken the lead in encouraging the recognition of digital scholarship in promotion and tenure cases. Its 2006 *Report of the MLA Task Force on Evaluating Scholarship for Tenure and Promotion* (`http://www.mla.org/tenure_promotion`) offered unequivocal support of digital scholarship. Among its recommendations are the following: "The profession as a whole should develop a more capacious conception of scholarship by rethinking the dominance of the monograph, promoting the scholarly essay, establishing multiple pathways to tenure, and using scholarly portfolios . . . .[and] departments and institutions should recognize the legitimacy of scholarship produced in new media, whether by individuals or in collaboration, and create procedures for evaluating these forms of scholarship.[2]"

This report and others (see "Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences" etc.) make it clear that the Humanities must broaden traditional definitions of scholarship and reconceptualize its methods of evaluation. But it still falls to us in the digital humanities to articulate the scholarly content and value of our work and to propose explicit procedures for effectively evaluating it.

In 2008 the MLA's Committee on Information Technology released an Evaluation Wiki[3] (based on the work of Geoffrey Rockwell who served as a member of the Committee from 2005-08) that provides a framework for departments to evaluate this new scholarship. As a result of this work, the authors of this paper led a workshop in evaluating digital work for tenure and promotion at the 2009 MLA convention in Philadelphia. The workshop was designed to provide both a framework from within which digital scholarship could be evaluated, as well as a forum for the authors to evaluate just how difficult it was for non-specialists to come to terms this new work. With this in mind, several themes were addressed in the workshop design:

1. The impediments involved in evaluating digital scholarship, particularly in terms of interdisciplinarity, authority, and use of standards;
2. Issues of disciplinarity: digital scholarship does not fit comfortably into the what is currently valued in the profession;
3. The absence of expertise at the departmental level in evaluating digital scholarship handicaps both the evaluators and the candidates;
4. The intellectual stakes of such work;
5. The value and credit accorded to interdisciplinary, multi-institutional, collaborative work.

In the field of literary studies the gold standard for tenure and promotion is the publication of the monograph. Editorial scholarship -- textual criticism, book history, scholarly editing, the sociology and bibliography of texts --particularly in North America, has been consistently devalued since the rise of literary theory. That the digital is conducive to the kinds of projects that have been denigrated by the academy (certainly since new literary criticism), including pedagogy, public humanities, and the creation of scholarly editions, has made the argument for including this work in tenure and promotion cases all the more difficult. These aforementioned themes and questions were meant to engage the group in questioning such values.

The workshop was structured around case studies based on the experiences of individuals in the field. In some instances the case studies were anonymized (as several of the workshop leaders created studies based on their own scholarship), others were not (as when it was clear who the author was from the online scholarship under discussion).

Attendees were seated at tables of eight and for each of three rounds had two case studies to choose from. The first round of case studies were what might term fairly traditional digital scholarly editions. This type of digital scholarship was chosen because it most resembled analogue scholarship. Nevertheless, when we opened the conversation up for discussion, the opening remark by a Departmental Chair questioned that the edition under discussion should be counted towards scholarship at all, as in her opinion, the creation of scholarly editions fit squarely into the category of service. This became a recurring theme of the workshop: individuals who did not understand the theoretical and technical imperatives behind the scholarship consistently categorized the work as service. It became clear that the production of editions, whether distributed in analogue or digital form, were not viewed as scholarship on par with critical theory.

Moreover, when the authors of case studies explained the theoretical methodology implicit in their research in terms of the technologies used, non-technical participants consistently complained that the authors were engaging in techno-speak jargon. Participants also complained that the work was undertheorized: however, when it was pointed out that theories common to digital scholarship (for example, the limitations of particular metadata schemes, issues arising from the hierarchical nature of XML, or the impediments to using a standard such as TEI for genetic editing) were evoked and challenged, these participants felt that it was not sufficient to rely on theoretical perspectives not normative to the discipline of literature (i.e. coding was not satisfactory as a basis for the field).

Subsequent case studies teased out issues of the implications of, in effect, outsourcing the evaluation of scholarship to (scholarly) presses and journals, the vast majority of which have not developed an economic model for publishing digital scholarship. Discussions circled around alternative models of validating digital scholarship, focusing on the work, for example, of NINES. Indeed, the success of the NINES model was so apparent in these discussions that NINES has committed its next two Summer Institutes (generously funded by the National Endowment for the Humanities) to evaluating digital scholarship.

Although the writing is on the wall that our field must change to encompass new scholarship and new scholarly forms as engendered by the new technologies, the impediments to evaluating this scholarship by non-specialists are paramount. The digital archive may have been the first form to challenge the primacy of print scholarship, but new forms of scholarship are emerging that will even more radically challenge the status quo.

It is clear from the research carried out by the authors in the preparation and delivery of this workshop, the creation of the CIT Guidelines, and the development of the NINES Summer Institutes, that it is imperative for the Digital Humanities community to take a more proactive role in supporting departments as they are increasingly called upon to evaluate digital scholarship. This burden cannot continue to be borne solely by each individual candidate, as is currently the norm. The overriding question for these activities must be how do we provide a framework so that evaluators can evaluate the research within the technological context within which the scholarship is undertaken.

The discussion of the case studies by the participating department chairs, senior faculty members, and junior faculty members revealed to us some of the specific difficulties departments and institutions have in recognizing scholarly activity in some of its new digital forms:

1. The evolving definitions of scholarship in language and literature over the past 50 years (Eagelton [4]), in particular the conflict between criticism and philology as the dominant mode of scholarship;
2. The discounting of scholarly activities like textual editing and translation, and the mislabeling of much digital scholarship as service[5] (McGann,[6] Gabler[7]);
3. The ongoing crisis in scholarly book publishing (Waters,[8] Guillory,[9] Greenblatt[10])

Our paper will summarize what happened at the workshop, in response to which we will present action items to be undertaken by organizations such as MLA and ACH.

Notes

1. For just a small sampling of recent postings and articles see Scott Jaschik, 'Tenure in a Digital Era'. Inside Higher Education (26 May 2010); 'Evaluating Digital Scholarship, Promotion & Tenure Cases', Office of the Dean, University of Virginia, http://artsandsciences.virginia.edu/dean/facultyemployment/evaluating_digital_scholarship.html; 'Guidelines for Evaluating Work with Digital Media in the Modern Languages', Modern Language Association, http://www.mla.org/guidelines_evaluation_digital; Julia Flanders, 'The Productive Unease of 21st-century Digital Scholarship'. Digital Humanities Quarterly (Summer 2009); Jeanne Glaubitz Cross, 'Reviewing Digital Scholarship: The Need for Discipline-Based Peer Review'. Journal of Web Librarianship (December 2008); Joan F. Cheverie, et al, 'Digital Scholarship in the University Tenure and Promotion Process: A Report on the Sixth Scholarly Communication Symposium at Georgetown University Library' Journal of Scholarly Publishing (April 2009) 219-230.

2. Report of the MLA Task Force on Evaluating Scholarship for Tenure and Promotion. http://www.mla.org/tenure_promotion
3. The Evaluation of Digital Work. http://wiki.mla.org/index.php/Evaluation_Wiki
4. Terry Eagelton. "The Functions of Criticism," How To Read a Poem (Malden, MA: Basil Blackwell 2007), 1-24.
5. Connie Moon Sehat and Erika Farr. 'The Future of Digital Scholarship: Preparation, Training, Curricula Report of a colloquium on education in digital scholarship'. (2009). http://www.clir.org/pubs/resources/articles.html
6. McGann, Jerome. "A Note on the Current State of Humanities Scholarship." Critical Inquiry 30/2. http://criticalinquiry.uchicago.edu/issues/v30/30n2.McGann.html
7. Gabler, Hans Walter. "Theorizing the Digital Scholarly Edition." Literature Compass 7/2 (2010) 43-56.
8. Lidnsay Waters. "A Modest Proposal for Preventing the Books of the Members of the MLA from Being a Burden to Their Authors, Publishers, or Audiences." PMLA 115 (2000): 315–17.
9. Guillory, John. "Evaluating Scholarship in the Humanities: Principles and Procedures." ADE Bulletin 137 (Spring 2005), pp. 18–33.
10. Greenblatt, Stephen. "A Special Letter." 28 May 2002. 1 Nov. 2010 http://www.mla.org/scholarly_pub

Automatic Extraction of Catalog Data from Genizah Fragments' Images

Shweka, Roni
rshweka@genizah.org
The Friedberg Genizah Project, Israel; Tel Aviv University, Israel

Choueka, Yaacov
ychoueka@genizah.org
The Friedberg Genizah Project, Israel; Tel Aviv University, Israel

Wolf, Lior
liorwolf@gmail.com
The Friedberg Genizah Project, Israel; Tel Aviv University, Israel

Dershowitz, Nachum
nachumd@post.tau.ac.il
The Friedberg Genizah Project, Israel; Tel Aviv University, Israel

Zeldin, Masha
mzeldin@genizah.org
The Friedberg Genizah Project, Israel; Tel Aviv University, Israel

The Cairo Genizah is a collection of about 250,000 manuscript-fragments of mainly Jewish texts discovered in the late 19th century. Most of the fragments were written between the 10th and the 14th centuries. Today, the fragments are spread out in libraries and private collections worldwide. The Friedberg Genizah Project (www.genizah.org) is in the midst of a multi-year process of digitally photographing all of the extant fragments. As of March 2011, the virtual library of the project holds over 250,000 digital images, and 200,000 additional images are expected to be integrated over the next few years. Unfortunately, this huge collection is far from being entirely cataloged, despite the ongoing effort to document and catalog all extant fragments. Moreover, the existing catalogs differ greatly in the amount and type of the data they present. Many of them record briefly the content of the fragment, without any information regarding its physical attributes.

We present a system for collecting all catalog data that can be extracted automatically from the fragment's image, mainly: the exact dimensions of the fragment; number of columns; number of lines; size of the margins; the fragment's physical status (torn vertically, horizontally, missing corners); and several additional features. Our system differentiates between bifolios and single pages, and in the first case collects the above data for each page separately. Besides the above attributes, which are expected to be found in every modern catalog, the system extracts some finer data that may be relevant to paleographic studies, such as density of lines (line height, inter-line space) and density of characters (number of characters in a fixed unit of length).

In addition to the detailed physical description of a single fragment, the huge database generated by the system serves for supporting identification of "join" candidates in the Cairo Genizah. A *join* is a set of manuscript-fragments that originate from the same original codex, but are scattered today under different shelfmarks, possibly in several different libraries. In a previous work, we described a system for the automatic identification of joins by ascertaining the degree of handwriting similarity between pairs of fragments. By querying the database and applying some basic rules for a good match, taking into account the completeness or incompleteness of the fragments, we can significantly improve on the quality of the results obtained by just analyzing the handwriting similarity.

Another aspect introduced in this paper is the proper conditions for taking digital images of manuscripts that are necessary for achieving this kind of results. We argue that, today, the function of such digital imaging is not only conservation and accessibility, but these images should be considered as potential inputs to image-processing algorithms and processes, and the computer should be therefore taken into account as one of the "clients" of the images. Hence, appropriate conditions should be considered in advance when digitizing manuscripts. Among these conditions we mention the following:

- **Choosing the optimal background for foreground-background separation.** The background color should contrast, not only with the color of the fragment material, (vellum or paper), which is some hue of light brown, but also with the color of the ink, usually dark brown or black. Otherwise, text will be erroneously recognized as part of the background, and characters will be interpreted as holes in the fragment. The common practice in some libraries to digitize manuscripts on white, brown or black background should be considered therefore as an imperfect one, because these colors do not contrast well with the manuscript

and the ink colors. Our study shows that the best contrast for these colors is provided by a blue background. Indeed, when we started digitizing the huge Genizah collection at the Cambridge University Library, we used blue as the standard background color for all images and the same practice was followed in the digitization of the British Library Genizah collection. Note that since with such contrasting colors the computer can very effectively differentiate between the fragment and its background, it is possible to automatically change the color background to any color desired by the user.

- **Avoiding the use of clips, weight bags, notes, etc.** Every significant element in the image should be easily identified and recognized by the computer, and the best segmentation is achieved by color separation. On the other hand, when there is a need for use of extra elements with no significance to appear, such as elements to hold the fragment or keep it flat, we recommend that they be of the same color as that of the background. Notes (such as shelfmark numbers) should be of a fixed size and shape, with some apparent icon on them, so as to enable the software to identify them easily.
- **Use of a ruler in the image.** Placing a ruler in the image enables the software to automatically determine the exact dpi of the image, and thus assess the various measures in some recognized unit, such as cms or inches. This practice is crucial especially when different images are taken with different lenses or when the camera is not fixed in the same position throughout the entire process. The ruler should be distinctive from the fragment; hence a wooden brown ruler or a see-through plastic one will not make a good choice.

Unfortunately, when such aspects are neglected, the application of computerized methods as described above and harvesting their results become unnecessarily difficult, and the quality of obtained results is adversely affected.

References

Lerner, HG & Jerchower, S (2006). 'The Penn/ Cambridge Genizah fragment project: issues in description, access, and reunification'. *Cataloging & Classification Quarterly.* 1: 21-39.

Reif, SC (2000). *A Jewish archive from Old Cairo: the history of Cambridge University's Genizah collection.* Richmond: Curzon Press.

Stinson, T (2009). 'Codicological Descriptions in the Digital Age'. *Codicology and Palaeography in the Digital Age - Kodikologie und Paläographie im Digitalen Zeitalter.* M Rehbein, P Sahle & T Schaßan (eds.) (ed.). BoD, Norderstedt: Schriftenreihe des Instituts für Dokumentologie und Editorik. V. 2, pp. 35-51.

TEI Consortium (2011). 'Manuscript Description'. *TEI P5: Guidelines for electronic Text Encoding and Interchange, Version 1.9.1.* `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html` (accessed 15 March 2011).

The Friedberg Genizah Project. `http://www.genizah.org/`.

Wolf, L, Littman, R, German, T, Mayer, N, Dershowitz, N, Shweka, R & Choueka, Y (2011). 'Automatically identifying join candidates in the Cairo Genizah'. *International Journal of Computer Vision.* `http://www.springerlink.com/content/p227026r1124xj30/fulltext.pdf`.

A Trip Around the World: Balancing Geographical Diversity in Academic Research Teams

Siemens, Lynne
siemensl@uvic.ca
School of Public Administration, University of Victoria

Burr, Elisabeth
elisabeth.burr@uni-leipzig.de
Institute of Romance Studies, University of Leipzig

Cunningham, Richard
richard.cunningham@acadiau.ca
Acadia Digital Culture Observatory, Acadia University

Duff, Wendy
wendy.duff@utoronto.ca
Faculty of Information, University of Toronto

Forest, Dominic
dominic.forest@umontreal.ca
École de bibliothéconomie et des sciences de l'information, Université de Montréal

Warwick, Claire
c.warwick@ucl.ac.uk
UCL Centre for Digital Humanities, Department of Information Studies, University College London

1. Introduction

Interdisciplinary and multidisciplinary projects are becoming an important part of academic life. Research questions are becoming more complex and sophisticated, requiring a team approach to address (Newell et al., 2000, Hara et al., 2003). At the same time, funding agencies are also encouraging these types of projects through their granting programs. One result of this increased level of research collaboration is that teams have members located at other institutions, whether nationally or internationally. Digital Humanities as a community is becoming increasingly international in focus. For example, the Digging into Data Challenge was jointly sponsored by Canada's Social Sciences and Humanities Research Council (SSHRC), the British Joint Information Systems Committee (JISC) and the American National Endowment for the Humanities (NEH) and National Science Foundation (NSF) and required each research team to have membership from at least two of the participating countries (Office of Digital Humanities, 2010). Further, the Digital Humanities Summer Institute at the University of Victoria has had participants from every continent, except the Antarctic. Finally, the Text Encoding Initiative Consortium has members and users from around the world, with a growing number of projects based in Asia (Siemens, 2008c). These collaborations are possible given the advances in computers and telecommunications. The recruitment of the right person is no longer limited by geography (Cramton et al., 2005).

However, these types of teams encounter the general challenges associated with collaborative work, but also with more specific ones relating to the geographical distribution of members. Despite this increasing use of research teams, protocols to prepare individual researchers to work as part of a team, particularly within those groups which cross language, culture, and country lines, are not widely developed. Our work is designed to identify effective work patterns and intra-team relationships and describe the support and research preparation that is required to develop international research collaborations. The results will enable those who work in such teams to recognize those factors that tend to predispose them to success, and perhaps more importantly, to avoid those that may lead to problematic interactions, and thus make the project less successful than it might have otherwise been.

2. Context

While there is a growing knowledge base of the benefits and challenges inherent in academic research teams (See, for example: Bracken et al., 2006, Choi et al., 2007, Fiore, 2008, Kraut et al., 1987), little knowledge exists about the ways in which teams with memberships across universities and disciplines work together (Garland et al., 2006), much less about teams with members from multiple country, culture and language groups (Setlock et al., 2004, Shore et al., 2005).

At a practical level, geography presents relatively simple challenges. As time zones increase, the flexibility in scheduling meetings decreases while the cost of face-to-face interactions increases. Technology may not always be capable of overcoming these challenges. Distance between team members limits the amount of interaction needed for creativity and innovation (Cummings et al., 2005, Lawrence, 2006). Further, some technologies and software may not be available in some countries due to infrastructure gaps

or government policy (for example, the Great Firewall of China).

At a more complex level, differences in culture and language may impact on various aspects of team work such as structure, management, communication, conflict expression and resolution, decision making, and appropriate team behaviour and may be further complicated by professional and academic cultural differences (Dafoulas, 2002, Setlock et al., 2004, Shachaf et al., 2007, Fry et al., 2007, Lee-Kelley et al., 2008, Dekker et al., 2008). At a very basic level, teams must decide a working language, a decision that may be political in nature (Butler, 1998, Deepwell et al., 2009, Bournois et al., 1998). And even with a common language, members may find that they must still translate terms. For example, institutions in different countries define a research assistant (RA) in a variety of ways. In Canada, an RA is generally a graduate student who works on a research project on a part-time basis while in the United Kingdom, an RA is a post-doctoral fellow who is on a full-time contract for a specified period of time. As a result, confusion can occur among team members when they use common terms in different ways.

Research suggests several factors that may minimize the impact of the above challenges. First, education and training may mitigate the impact of cultural differences because team members may share professional norms. University education is fairly similar across countries (Dafoulas, 2002, Lee-Kelley et al., 2008, Nason et al., 1998). Teams may also find it beneficial to spend time in members' cultures to create understanding of differences and similarities (Nason et al., 1998, Bagshaw et al., 2007). Finally, teams may also consider creating a cultural profile, both by country and professional/ discipline culture (Dafoulas, 2002, Zakaria et al., 2008). This profile can be combined with team norms to express understandings of time, deadlines, language, communication channels, conflict resolution mechanisms, and other issues (Saunders et al., 2004). However, more research is needed to understand the supports and research preparation that is needed to ensure effective collaborations in teams with memberships from many countries, cultures, and language groups.

3. Methodology

This paper is part of a larger project examining research teams with multi-country, culture and language representation, led by a team based in Canada, United Kingdom and Germany (For more details, see Siemens, 2010). The larger study uses a combination of data collection methods including an ethnomethodological approach with diary/log studies of research teams in the midst of their collaboration (Garfinkel, 1984) and semi-structured interviews with individuals who have experiences in the types of teams under investigation (Marshall et al., 1999). This paper will report on the findings from the interviews.

Digital Humanities community will serve as an exemplar case study population given the community's international focus and collaborative networks. To achieve their objectives, and because of the variety of skills and expertise that these projects require, DH researchers must work collaboratively both within their institutions and with others nationally and internationally and often across disciplines. Team members include humanities, social science and computer science scholars, undergraduate and graduate students, research assistants, computer programmers and developers, librarians, and others. At present, several research projects involving national and international teams with funding ranging from thousands to millions of dollars are already in place with others in development. In addition, several DH research centres have faculty and research staff drawn from several countries, creating a mix of languages and cultures.

This research project builds on earlier work on DH research teams presented at previous digital humanities conferences (Siemens, 2008a, Siemens, 2008b, Siemens et al., 2009a, Siemens et al., 2010).

4. Preliminary Findings

At the time of writing this proposal, interviews are being conducted and data analysis completed.

The benefits to the Digital Humanities community will be several. First, the study contributes to an explicit description of the community's work patterns and relationships, particularly as the Digital Humanities community continues to become international in focus (Alliance of Digital Humanities Organizations, 2010, Tei-C, 2010). This research demonstrates that much of digital humanities research is accomplished within interdisciplinary research teams, which are developing tools and processes to facilitate this collaboration. One particular issue highlighted in this research relates to challenges experienced within teams with members from various countries and cultures (Siemens et al., 2009b). Second, it is designed to enable those who work in such teams to recognise factors that tend to predispose them to success, and perhaps more importantly, to avoid those that may lead to problematic

interactions, and thus make the project less successful than it might otherwise have been.

References

Alliance of Digital Humanities Organizations (2010). About.. http://digitalhumanities.org/about (accessed accessed September 26, 2010).

Bagshaw, D., Lepp, M. & Zorn, C. R. (2007). 'International Research Collaboration: Building Teams and Managing Conflicts.'. *Conflict Resolution Quarterly.* Vol 24, No 4: 433-446.

Bournois, F. & Chevalier, F. (1998). 'Doing research with foreign colleagues: a project-life cycle approach.'. *Journal of Managerial Psychology.* Vol 13, No 3/4: 206-213.

Bracken, L. J. & Oughton, E. A. (2006). '"What do you mean?' The importance of language in developing interdisciplinary research.'. *Transactions of the Institute of British Geographers.* Vol 31, No 3: 371-382.

Butler, M. C. (1998). 'An integrating perspective on interdisciplinary, cross-cultural research.'. *Journal of Managerial Psychology.* Vol 13, No 3/4: 199-205.

Choi, B. C. K. & Pak, A. W. P. (2007). 'Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 2. Promotors, barriers, and strategies of enhancement.'. *Clinical & Investigative Medicine.* Vol 30, No: E224-E232.

Cramton, C. D. & Webber, S. S. (2005). 'Relationships among Geographic Dispersion, Team Processes, and Effectiveness in Software Development Work Teams.'. *Journal of Business Research.* Vol 58, No 6: 758-765.

Cummings, J. N. & Kiesler, S. (2005). 'Collaborative Research Across Disciplinary and Organizational Boundaries.'. *Social Studies of Science.* Vol 35, No 5: 703-722.

Dafoulas, G. (2002). 'Investigating Cultural Differences in Virtual Software Teams.'. *Electronic Journal on Information Systems in Developing Countries.* Vol 7, No: 1-14.

Deepwell, F. & King, V. (2009). 'E-Research Collaboration, Conflict and Compromise.'. *Handbook of Research on Electronic Collaboration and Organizational Synergy.* Salmons, J. & Wilson, L. (ed.). Hershey, Pennsylvania: IGI Global.

Dekker, D. M., Rutte, C. G. & Van Den Berg, P. T. (2008). 'Cultural differences in the perception of critical interaction behaviors in global virtual teams.'. *International Journal of Intercultural Relations.* Vol 32, No 5: 441-452.

Fiore, S. M. (2008). 'Interdisciplinarity as Teamwork: How the Science of Teams can Inform Team Science.'. *Small Group Research.* Vol 39, No 3: 251-277.

Fry, J. & Talja, S. (2007). 'The intellectual and social organization of academic fields and the shaping of digital resources.'. *Journal of Information Science.* Vol 33, No 2: 115-133.

Garfinkel, H. (1984). *Studies in Ethnomethodology.* Malden, MA: Polity Press/Blackwell Publishing.

Garland, D. R., O'connor, M. K., Wolfer, T. A., et al. (2006). 'Team-based Research: Notes from the Field.'. *Qualitative Social Work.* Vol 5, No 1: 93-109.

Hara, N., Solomon, P., Kim, S.-L., et al. (2003). 'An Emerging View of Scientific Collaboration: Scientists' Perspectives on Collaboration and Factors that Impact Collaboration.'. *Journal of the American Society for Information Science and Technology.* Vol 54, No 10: 952-965.

Kraut, R. E., Galegher, J. & Egido, C. (1987). 'Relationships and Tasks in Scientific Research Collaboration.'. *Human-Computer Interaction.* Vol 3, No 1: 31-58.

Lawrence, K. A. (2006). 'Walking the Tightrope: The Balancing Acts of a Large e-Research Project.'. *Computer Supported Cooperative Work: The Journal of Collaborative Computing.* Vol 15, No 4: 385-411.

Lee-Kelley, L. & Sankey, T. (2008). 'Global virtual teams for value creation and project success: A case study.'. *International Journal of Project Management.* Vol 26, No 1: 51-62.

Marshall, C. & Rossman, G. B. (1999). *Designing Qualitative Research.* Thousand Oaks, CA: SAGE Publications.

Nason, S. W. & Pillutla, M. M. (1998). 'Towards a Model of International Research Teams.'. *Journal of Managerial Psychology.* Vol 13, No 3/4: 156-166.

Newell, S. & Swan, J. (2000). 'Trust and Inter-organizational Networking.'. *Human Relations.* Vol 53, No 10: 1287-1328.

Office of Digital Humanities (2010). *Digging into Data Challenge..* http://www.diggingintodata.org/ (accessed accessed September 26 2010).

Saunders, C., Van Slyke, C. & Vogel, D. R. (2004). 'My time or yours? Managing time visions in global virtual teams.'. *Academy of Management Executive.* Vol 18, No 1: 19-31.

Setlock, L. D., Fussell, S. R. & Neuwirth, C. (2004). 'Taking it out of context: collaborating within and across cultures in face-to-face settings and via instant messaging.'. *Proceedings of the 2004 ACM conference on Computer supported cooperative work.* Chicago, IL.

Shachaf, P. & Hara, N. (2007). 'Behavioral complexity theory of media selection: a proposed theory for global virtual teams.'. *Journal of Information Science.* Vol 33, No 1: 63-75.

Shore, B. & Cross, B. J. (2005). 'Exploring the role of national culture in the management of large-scale international science projects.'. *International Journal of Project Management.* Vol 23, No 1: 55-64.

Siemens, L. (2008). 'The Balance between On-line and In-person Interactions: Methods for the Development of Digital Humanities Collaboration.'. *SDH-SEMI 2008.* Vancouver, BC.

iemens, L. (2008). '"It's a team if you use 'reply all'": An Exploration of Research Teams in Digital Humanities Environments.'. *Digital Humanities 2008.* Oulu, Finland.

Siemens, L. (2008). 'Understanding the TEI-C Community: A Study in Breadth and Depth, Toward Membership and Recruitment.'. *TEI Annual Conference.* London, UK.

Siemens, L. (2010). 'Understanding Academic Research Teams: Implications of Multi-Country, Multi-Language, and Multi-Culture Team Membership.'. *European Summer School.* Leipzig, Germany.

Siemens, L., Cunningham, R., Duff, W., et al. (2009). '"More minds are brought to bear on a problem": Methods of Interaction and Collaboration within Digital Humanities Research Teams.'. *Society for Digital Humanities/Société pour l'étude des médias interactifs.* Ottawa, Ontario.

Siemens, L., Cunningham, R., Duff, W., et al. (2010). 'A Tale of Two Cities: Implications of the Similarities and Differences in Collaborative Approaches within the Digital Libraries and Digital Humanities Communities.'. *Digital Humanities 2010.* London, UK.

Siemens, L., Duff, W., Warwick, C., et al. (2009). '"It challenges members to think of their work through another kind of specialist's eyes": Exploration of the Benefits and Challenges of Diversity in Digital Project Teams. Thriving on Diversity - Information Opportunities in a Pluralistic World'. *ASIS&T 2009 Annual Meeting.* Vancouver, British Columbia.

Tei-C (2010). *Current TEI Members..* `http://www.tei-c.org/Membership/current.xml` (accessed accessed October 12, 2010).

Zakaria, N., Amelinckx, A. & Wilemon, D. (2008). 'Navigating Across Culture and Distance: Understanding the Determinants of Global Virtual Team Performance.'. *Work Group Learning: Understanding, Improving & Assessing How Groups Learn in Organizations.* Sessa, V. I. & London, M. (ed.). New York, NY: Lawrence Erlbaum Associates.

Mining Language Resources from Institutional Repositories

Simons, Gary F.
gary_simons@sil.org
SIL International and Graduate Institute of Applied Linguistics

Bird, Steven
sb@csse.unimelb.edu.au
University of Melbourne and University of Pennsylvania

Hirt, Christopher
chris_hirt@sil.org
SIL International and Payap University

Hou, Joshua
jshou@u.washington.edu
University of Washington

Pedersen, Sven
sven.pedersen@gmail.com
Graduate Institute of Applied Linguistics

Language resources are the bread and butter of language documentation and linguistic investigation. They include the primary objects of study such as texts and recordings, the outputs of research such as dictionaries and grammars, and the enabling technologies such as software tools and interchange standards. Increasingly, these resources are maintained in digital form and distributed via the web. However, searching on the web for language resources is a hit-and-miss affair. One problem is that many online resources are hidden behind interfaces to databases with the result that only a fraction of these resources are being indexed by search engines (He and others 2007). Even when resources are exposed to online search engines, they may not be discoverable since they are described in ad hoc ways that prevent searches from retrieving the desired results with high recall or precision.

This paper describes work being done in the context of the Open Language Archives Community (OLAC) to develop a service that uses text mining methods (Weiss and others 2005) to find language resources located within the hidden web of institutional repositories. It then uses the OLAC infrastructure to expose them on the open web and make them discoverable through precise search.

2. The OLAC Infrastructure

As set out in its mission statement, the Open Language Archives Community[1] is "an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources."

With respect to best practices, the community has thus far focused on developing recommendations for the metadata description of language resources[2] so that they can be discovered in search with high precision and recall. The OLAC metadata format[3] is an extension of Dublin Core[4]—the dominant metadata standard in the digital library and World Wide Web communities (Bird and Simons 2004). To support the need for precise search, the community has adopted five specialized vocabularies[5] for use in describing resources: *subject language*, for identifying precisely which language(s) a resource is "about" by using a code from ISO 639;[6] *linguistic type*, for classifying the structure of a resource as primary text, lexicon, or language description; *linguistic field*, for specifying relevant subfields of linguistics; *discourse type*, for indicating the linguistic genre of the material; and *role*, for documenting the parts played by specific individuals and institutions in creating a resource.

With respect to the network of interoperating repositories, there are now more than 40 institutions that are sharing their language resource metadata to create a virtual digital library with over 90,000 holdings. Participating archives publish their catalogs in the XML format of an OLAC repository[7] and these repositories are "harvested" thrice daily by the OLAC aggregator using the Open Archives Initiative (OAI) Protocol for Metadata Harvesting[8] (Simons and Bird 2003)—another standard of the digital library community.

3. Mining for Hidden Language Resources

The Open Access movement[9] has led to the widespread uptake of self-archiving of research results by university faculty and staff. It stands to reason that among the millions of resources deposited into open-access institutional repositories, there are thousands of language resources. But these resources are not typically accessible via general web search. This is because they are hidden behind the search interfaces of hundreds of repositories and they lack precise identification as language resources. The question is,

"Can we find the language resources in institutional repositories and then make them easy for the language resources community to discover?"

Our research addresses the problem by using text mining techniques. We have begun by training a binary classifier that identifies the likely language resources within an institutional repository. We used MALLET, the Machine Learning for Language Toolkit,[10] to train a maximum entropy classifier. For data to train the classifier we needed a large collection of metadata records covering the full range of human knowledge that were already classified as to the nature of their content. For this purpose we used the collection of more than 9 million MARC catalog records from the Library of Congress collection that was deposited into the Internet Archive[11] by the Scriblio project.[12] We used bag-of-words features extracted from the title and subject headings of each MARC record. To label each record as to whether it was a language resource or not, we mapped the Library of Congress call number onto the appropriate binary label based on a prior analysis of the Library of Congress classification system. The resulting set of 9 million training records was then given to MALLET to train a binary classifier for language resource identification.

The resulting classifier was applied to over 4 million Dublin Core metadata records that were collected by doing a complete harvest of nearly 700 institutional repositories using the OAI Protocol for Metadata Harvesting. The list of base URLs to harvest was found by going to the University of Illinois OAI-PMH Data Provider Registry[13] and querying for all repositories with the word "university" in their Identify response. When applied to a metadata record, the classifier returns a number between 0 and 1 representing the probability that the resource is a language resource. This probability was added as a new metadata element to each harvested record. We then implemented an extension to the OAI-PMH interface on our metadata aggregator that allows us to request a ListRecords response of a given size that is a random sample of the records falling within a given probability range.

Figure 1 shows the result of evaluating the performance of the classifier by means of manually inspecting ten random samples of 100 records each representing the full range of probabilities assigned by the classifier. In the manual evaluation of the classifier results, each record was assigned to one of three categories: not a language resource, a resource about a specific language, or a resource about human language but no language in particular. Figure 1 plots the number of specific language resources found in each sample of 100 as the lower line; the upper line adds the non-specific language resources. (Not plotted are a sample of 500 records for .001 < p < .01 in which were found 0 specific language resources and 4 non-specific resources, and a sample of 200 records for p < .001 in which 0 language resources of either type were found.) The graph demonstrates that the probabilities assigned by the classifier accord well with the actual proportions discovered by manual inspection, thus providing evidence for the validity of the classifier. The notable deviation from the expected trend is in the highest probability range. Inspection of the records in question showed that the majority of false positives were items from computer science about programming languages and formal language theory, leading us to hypothesize that the training data from the Library of Congress catalog was underrepresented in this area.

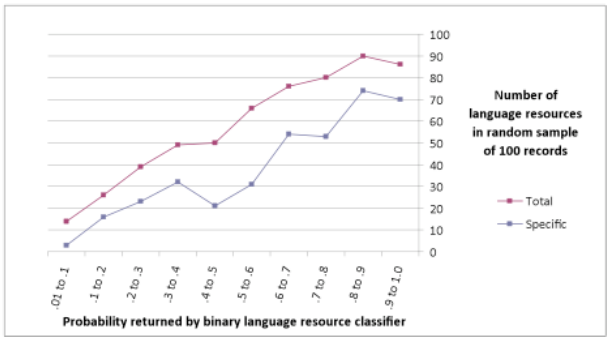


Figure 1: Evaluation of language resource classifier

Of the 4 million harvested records, only 52,000 indicate a probability greater than .01 of being a language resource. Multiplying the proportion of actual language resources found within each probability range by the total number of records falling within each range leads to the estimate that there are approximately 8,000 specific language resources within the set of 4 million harvested records.

4. Exposing the Once-Hidden Resources

The next step in our research is to apply a multiclass classifier for language resource types to the metadata records for the 52,000 candidate language resources, as well as a named entity recognizer for language names. The metadata records to which language resource type and language identification can be assigned with high probability will be enriched using the OLAC metadata vocabularies. They will then be entered into the combined OLAC catalog by creating a new OLAC data provider for these language resources that have been mined from institutional repositories. The final paper will report on the results of these efforts

at metadata enrichment and show how the results are exposed to users through the two main OLAC services that support language resource discovery: an indexing service that provides a web page of relevant resources for each of 7,670 distinct human languages (as identified in the ISO 639-3 standard) and a faceted search service that makes it easy to find resources of interest by clicking on selected values of standardized descriptors to successively refine the search.

References

Bird, Steven and Gary Simons (2004). 'Building an Open Language Archives Community on the DC Foundation'. *Metadata in Practice [D. I. Hillmann and E. L. Westbrooks, eds.].* Chicago: American Library Association.Pp. 203–222. http://www.ldc.upenn.edu/sb/home/papers/mip.pdf.

He, Bin, Mitesh Patel, Zhen Zhang, and Kevin Chen-Chuan Chang (2007). 'Accessing the deep web'. *Communications of the ACM.* 50(5): 95–101.

Simons, Gary and Steven Bird (2003). 'Building an Open Language Archives Community on the OAI Foundation'. *Library Hi Tech.* 21(2): 210–218. http://arxiv.org/abs/cs.CL/0302021.

Weiss, Sholom M., Nitin Indurkhya, Tong Zhang, and Fred J. Damerau (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information.* New York: Springer.

Notes

1. http://www.language-archives.org/
2. http://www.language-archives.org/REC/bpr.html
3. http://www.language-archives.org/OLAC/metadata.html
4. http://dublincore.org/documents/dcmi-terms/
5. http://www.language-archives.org/REC/olac-extensions.html
6. http://www.sil.org/iso639-3/
7. http://www.language-archives.org/OLAC/repositories.html
8. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
9. http://en.wikipedia.org/wiki/Open_Access_movement
10. http://mallet.cs.umass.edu/
11. http://www.archive.org/details/marc_records_scriblio_net
12. http://about.scriblio.net/
13. http://gita.grainger.uiuc.edu/registry/

Knowing and Doing: Understanding the Digital Humanities Curriculum

Spiro, Lisa
lisamspiro@gmail.com
Rice University

As the digital humanities have become more visible, attracting attention both from publications such as the *Chronicle of Higher Education* and from universities eager to develop emerging areas of research and teaching, commentators have been debating what qualifies as digital humanities and whether the community is sufficiently inclusive. In part, as Geoffrey Rockwell suggests, this debate reflects digital humanities' failure to provide multiple paths to entry. In the past, many entered the digital humanities by apprenticing with established practitioners, but such opportunities are not available to all (Rockwell 2010). New educational programs such as the MA in Digital Humanities at University College London and the proposed MA in Knowledge and Networks at Duke are being put forward to address these needs, joining established programs at Kings College London, University of Alberta, and other universities. As educational opportunities expand, the digital humanities (DH) community should examine how digital humanists are being trained (Hirsch & Timney 2010). Based on this knowledge, the community can create a more coordinated (though still flexible) approach to the DH curriculum that reflects its own commitment to openness, collaboration, interdisciplinarity and experimentation. Education is fundamental both to how the community comprehends itself and how it brings in new members. As Melissa Terras observes, curriculum "can be seen to define the field in the way the publication record cannot," serving as a "hidden history" that reveals what knowledge experts believe to be crucial and how that knowledge is transmitted (Terras 2005, p.1). By examining digital humanities education, we can participate in a concrete conversation about how to make the field both more inclusive and more targeted toward the core knowledge and skills that digital humanists need.

In order to understand how the digital humanities are taught at universities today, I will look at both curriculum and courses.[1] First, to get a broader perspective on the digital humanities *curriculum*, I will examine degree requirements and curriculum plans

for digital humanities undergraduate, masters, and Ph.D. programs. How are digital humanities degree programs structured? What courses are deemed necessary, and how are they sequenced? To what extent are projects and/or internships required? How do explicitly digital humanities programs compare to more traditional programs that include a digital humanities component? How are collaboration and interdisciplinarity inculcated? Case studies of several programs will be offered to elucidate different approaches.

I am also analyzing a collection of over 200 DH *syllabi*[2] for both graduate and undergraduate courses. These syllabi represent a variety of approaches to digital humanities, such as media studies, text encoding, programming, and information visualization, and come from a range of departments, including history, English, digital humanities, library and information science, and computer science. I am examining:

- What do these syllabi say about how knowledge in the digital humanities is categorized and conceptualized?
- What are the course goals? How is learning assessed?
- How do these classes balance theory and practice?
- To what extent do the courses reflect "digital pedagogy," the thoughtful use of technology to foster learning?
- To what extent are blogging, Twitter, or other forms of networked communication part of the course?
- What are the most frequently assigned readings? Is a DH "canon" emerging?
- What are some typical—and atypical—assignments? To what extent are projects required? How are projects structured and evaluated?
- How does the course reflect the practices and norms of its departmental home? For example, how do courses on digital history compare to those on digital literary studies or digital media studies?
- How is humanistic as well as technical knowledge woven into the course? What scholarly values do these courses promote?

I plan to use a combination of manual and automated methods to analyze the syllabi corpus. I am tagging the syllabi using a set of custom keywords so that I can sort them easily and see emerging patterns. I am also compiling a linked bibliography of readings assigned in DH courses. In addition, I will experiment with methods such as topic modeling and word frequency analysis to categorize the syllabi and extract key concepts.

Although my research is still underway, my initial analysis of a subset of approximately 50 syllabi suggests that:

- Digital humanities courses tend to use readings that are freely available online, such as *A Companion to Digital Humanities* and *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*, as well as blog posts by authors such as Dan Cohen and Bill Turkel. Given its apparent interest in up-to-date, openly accessible information, the community would benefit from openly sharing educational resources such as syllabi, assignments, exercises, and tutorials.
- While many digital humanities courses do focus on text, as Melissa Terras noted in 2006 (Terras 2006), they have expanded to include topics such as geospatial scholarship, multimedia design, and information visualization.
- Many courses require students to create scholarly digital projects, whether as individuals or in teams.

In a sense, this research updates Melissa Terras' 2005 study of the curriculum in humanities computing, charting how the field has changed in the last five years. I am returning to Terras' question about whether the digital humanities curriculum reflects the field's research agenda and whether DH has emerged as an " academic subject," capable of standing on its own (Terras 2005). I also hope my research helps to inform the digital humanities curriculum going forward. While the digital humanities curriculum should be nimble enough to keep up with the pace of technological change, diverse enough to encompass the variety of approaches to "big tent" digital humanities, and flexible enough to reflect the strengths of particular institutions, the community can benefit from greater coordination and sharing of educational resources to save labor, spread ideas, and provide greater coherence across programs. Of course, my analysis will focus on the characteristics of current DH education programs rather than on emerging areas of skill development, but I hope to document innovative educational approaches as well as points of consensus. Developing a DH curriculum may make it easier for departments (and individuals) to understand what they need to do to beef up their digital humanities portfolio and how they can specialize in particular areas of knowledge. Although my study will not define such a curriculum (which should be worked out by the community rather than an individual), I hope that it will illustrate trends and gaps in digital humanities education.

References

Hirsch, B. D., Timney, M (2010). 'The Importance of Pedagogy: Towards a Companion to Teaching Digital Humanities.'. *Digital Humanities 2010.* London, July 7-10, 2010, pp. 303.

Rockwell, G. (DATE). 'Inclusion In The Digital Humanities'. *philosophi.ca.* . http://www.philosophi.ca/pmwiki.php/Main/InclusionInTheDigitalHumanities (accessed June 22, 2010).

Terras, M. (2005). 'Disciplined: Using Curriculum Studies to Define 'Humanities Computing''. *ACH/ALLC 2005.* Victoria, BC, Canada, June 15 - June 18, 2005. http://pear.hcmc.uvic.ca:8080/ach/site/program.xq#abs_43.

Terras, M. (DATE). 'Disciplined: Using Educational Studies to Analyse 'Humanities Computing''. *Literary and Linguistic Computing.* 21(2): 229-246. http://llc.oxfordjournals.org/cgi/content/abstract/21/2/229.

Notes

1. I am in the process of collecting and synthesizing information related to this research, particularly through my Zotero group, “Digital Humanities Education” (http://www.zotero.org/groups/digital_humanities_education), and posts on my blog, “Digital Scholarship in the Humanities” (http://digitalscholarship.wordpress.com/). I am also working on a related project for Brett Hirsch’s collection *Teaching Digital Humanities: Principles, Practices, and Politics* proposing an open certificate for digital humanities.
2. All of these syllabi were found through web searches. Almost all are from universities in the United States, Canada and the UK. The earliest syllabus is from the late 1990s, but the bulk are from the late 2000s.

Layer upon Layer. “Computational Archaeology” in 15th Century Middle Dutch Historiography.

Stapel, Rombert
rstapel@fryske-akademy.nl
Fryske Akademy (KNAW) / Leiden University, Netherlands

A scholar or student who wishes to engage in ‘non-traditional’ authorship attribution would be wise to choose a test corpus that is as free as possible of ‘interfering’ features such as genre, external editing, or a corpus that is stretched over a large number of years. The higher the consistency throughout the corpus, the larger the chance of a successful investigation of authorship and/or stylistic features.

Medieval manuscripts are characterized by a much greater amount of, what we could call, ‘interfering’ features. Scribes manually copied texts again and again, not seldom altering the content and often altering the orthography. Most original works have been lost, just as much of the copied material. To add even more difficulties, what we nowadays will easily refer to as ‘original work’ is much less clear-cut in the Middle Ages. Our modern notion of ‘copyright’ is virtually unknown to medieval men and women and for a long time the concept of auctoritas (author) was primarily used in referral to classic writers such as Aristotle and Augustine. Many of the medieval texts are thus written anonymously.

The situation could be characterized as chaotic by scholars used to relatively straightforward text corpora. Before you can begin your quest for a medieval author, you first have to find out what content is scribal related and what can be attributed to the author. And just when you think you are making some progress, you find out that your ‘author’ has been merely compiling source texts, who he (or she) is copying word for word. A scholar addressing these texts should therefore meticulously peal of the different layers of the text. When confronted with these scenarios, it might not sound that surprising that the number of studies involving the use of computational techniques and medieval texts is not that great. In recent years though, some progress has been made (e.g. Van Dalen-Oskam and Van Zundert 2007; Van Dalen-Oskam,

Thaisen and Kestemont 2010). Most of all, these studies show that it is possible – using computational techniques such as Burrows' Delta – to overcome some of the difficulties in distinguishing for instance scribal and authorial layers within a single text.

2. The Case

In this paper I will contribute to this evolving field of research and discuss some possible methods that can be used to differentiate these different layers within medieval texts. Although the text corpus that I will be using originated from fully tagged TEI/XML files, for this purpose I will be using plain text files. This could be of great interest to scholars who are not able or not willing to spend large amounts of time, energy and skill in preparing their texts. The texts are transcribed without changing spelling to modern use (thus euen instead of even and Iherusalem instead of Jherusalem), but abbreviated words are expanded.

One of the more interesting aspects of this particular text corpus is that it has been written by a single person, whose name is recorded in one of his charters: Hendrik Gerardsz. van Vianen, most likely secretary of the Utrecht Land Commander of the Teutonic Order. The Teutonic Order was one of the three major military orders – beside the Knights Templars and Hospitallers – that defended the Holy Church in the Holy Land, the Iberian Peninsula and the Baltic region and received goods and land all over Europe, including the Low Countries. Hendrik van Vianen wrote at least 25 Middle Dutch charters containing land contracts for the Teutonic Order between 1479 and 1491. He also wrote a few Latin charters between 1489 and 1509 that are not included in this study. Furthermore, he manually copied a manuscript containing a Dutch version of the popular Sachsenspiegel that belonged to two Land Commanders of the Teutonic Order in Utrecht. Last but not least, he was responsible for a chronicle on the history of the Teutonic Order, known as the Croniken van der Duytscher Oirden or Jüngere Hochmeisterchronik (Stapel and Vollmann-Profe 2010). Codicological evidence suggests that the manuscript of his hand, now in Vienna, is an autograph, although it is not sure whether parts of the text existed before. Autographs are relatively rare in a medieval context, a recent survey of Middle Dutch manuscripts mentions barely more than a hundred examples (Houthuys 2009).

As a result, we have the unique opportunity to study a medieval text corpus – perhaps not as large as the ones modern literary scholars work with, but still substantial in medieval terms – of around 131.000 words that includes the work of a single person working as manuscript scribe; writer of land charters; and possible author of a history of his order.

Ever since the Croniken van der Duytscher Oirden has been studied in a scholarly context, there have been questions about the original composition. The original Middle Dutch version of the chronicle consists of a long prologue, a part that contains the deeds and lives of the Grand Masters of the Order, alternated by privileges granted by popes and emperors. It ends with a history of the regional bailiwick of Utrecht and its Land Commanders. Especially that last part is often put aside as a later addition to the chronicle. It is an interesting question, the more so since it has historical significance in the sense that it has an influence on the original function and aspiration of the text.

3. Methods

To examine the relationship between the different parts of the text corpus we have collected, I employed several methods, but for now, I would like to focus on the use of the Delta Spreadsheets, freely made available by David Hoover (Hoover 2009). Some primary samples were selected from of the text corpus using the Intelligent Archive (Craig, Whipp and Ralston 2010) to create word frequency lists: the combined charters; the Sachsenspiegel; the table of contents of the Croniken; its prologue; some of the privileges; three parts of the Grand Masters' part; and the bailiwick chronicle. These were computed against 2000 word pieces, overlapping and 500 words advancing. The results are shown in Figures 1, 2 and 3.

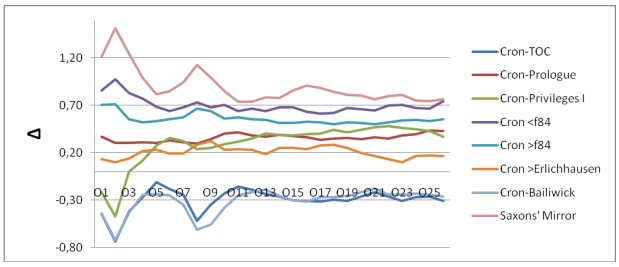


Figure 1: Moving Delta, Charters

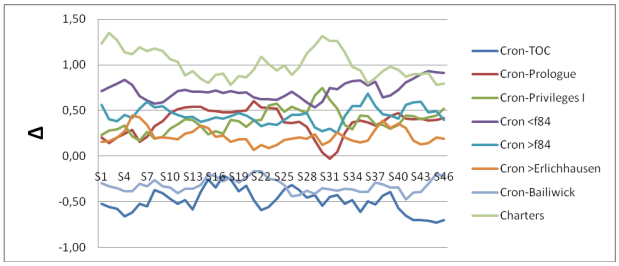


Figure 2: Moving Delta, Saxons' Mirror

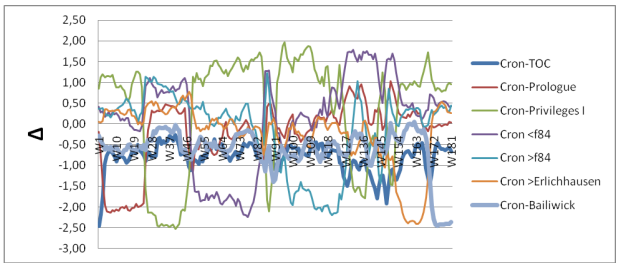


Figure 3: Moving Delta, Croniken

What stands out is that in all three instances, the primary samples of the table of contents and the bailiwick chronicle (in light and dark blue) outperform the other samples, especially if one leaves out the samples in Figure 3 which perform off course very well in their own consecutive areas.

4. Conclusion

What can be concluded from this observation? I think the table of content and the bailiwick chronicle best describe the personal writing style and orthography of Hendrik van Vianen. In both the charters, the copied Sachsenspiegel and the Croniken this layer is present, and it raises the question if he indeed can be held responsible for the bailiwick chronicle and the table of contents. The table of contents is clearly written at a later stage, added at the end of project, as is shown by watermark evidence and the distribution of some abbreviations and specific letter forms, quantified in Figure 4 and 5 below.

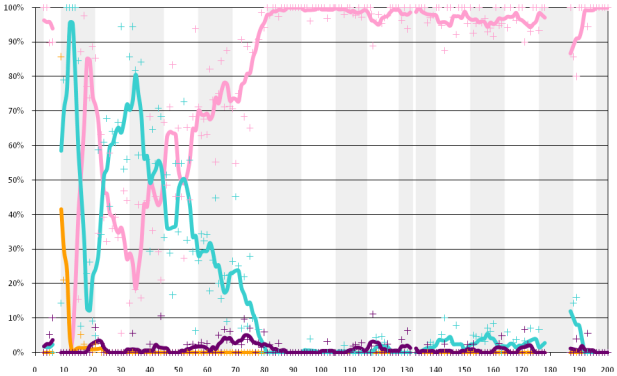


Figure 4: Different forms of letter “w”. The TOC in front corresponds with the latter part of the chronicle.

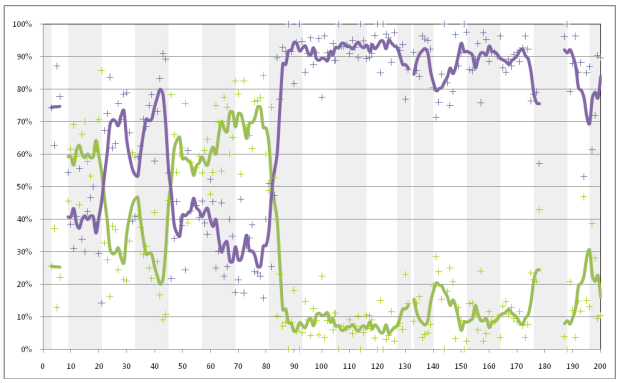


Figure 5: Use or absence of abbreviations in “ende” (Dutch for “and”).

Moreover, to what extent can Hendrik van Vianen still be held responsible for the rest of the chronicle? Was there an older exemplar of a chronicle available to him? It is hard to say for certain. There is a good possibility that he was not so much the – what we would now call – author, but more a compiler for large parts of the chronicle. The old source texts would then form another layer within the text. Again, in the Middle Ages it is not always possible (and even helpful) to make a strict distinction between the two. An old-fashioned approach to research the sources of the chronicles and how these source-texts were implemented could be a logical step further and will be one of the issues addressed in the remainder of this project.

References

Craig, H., Whipp, R., Ralston, M. (2010). *Intelligent Archive. Budgerigar Version.*. http://www.newcastle.edu.au/school/hss/research/groups/cllc/intelligent-archive.html.

Dalen-Oskam, K. van, Thaisen, J., Kestemont, M. (2010). 'Computational approaches to textual variation in medieval literature'. *Digital Humanties 2010 Conference Abstracts.* 37-44.

Dalen-Oskam, K. van, Zundert, J. van (2007). 'Delta for Middle Dutch – Author and Copyist Distinction in Walewein'. *Literary and Linguistic Computing.* 22: 345-362http://www.newcastle.edu.au/school/hss/research/groups/cllc/intelligent-archive.html.

Hoover, D. (2009). *The Excel Text-Analysis Pages.* http://https://files.nyu.edu/dh3/public/The%20Excel%20Text-Analysis%20Pages.html.

Houthuys, A. (2009). *De autograaf van de Brabantsche yeesten, boek VI (vijftiende eeuw). Hilversum 2009.. Middeleeuws kladwerk*

Kestemont, M., Dalen-Oskam, K. van (2009). 'Predicting the Past: Memory-Based Copyist and Author Discrimination in Medieval Epics'. *Proceedings of the Twenty-first Benelux Conference on Artificial Intelligence.* 121-128.

Stapel, R.J., Vollmann-Profe, G. (2010). 'Cronike van der Duytscher Oirden'. *Encyclopedia of the Medieval Chronicle.* Dunphy, G. (ed.). , pp. 328-329.

Thaisen, J. (2010). 'A Probabilistic Analysis of a Medieval English Text'. *Digitizing Medieval and Early Modern Material Culture.* Terras, M., Nelson, B. (eds.). , pp. 328-329.

Reforming Digital Historical Peer Review: Guidelines for Applying Digital Historiography to the Evaluative Process

Sternfeld, Joshua
jsternfeld@neh.gov
National Endowment for the Humanities, United States of America

From teaching students how to vet online websites, to formal peer review of digital publications, to evaluation of scholarship for tenure review, the need for a rigorous methodology to evaluate digital historical representations has never been more apparent. Sophisticated databases, digital libraries and archives, and virtual reconstructions have collectively reshaped engagement with the past that has challenged the boundaries of traditional historical practices. University professors encourage students to create mash-ups of historical multimedia content to be posted as YouTube documentaries. Museums employ mobile applications enriched with augmented reality to draw visitors further into an exhibition. Geospatial visualizations and virtual architectural reconstructions bring the past alive and challenge entrenched scholarly and popular perceptions.

While attention and resources have (justifiably) been focused on content creation and tool development, the digital history community has, until recently, neglected the development of a methodology that can evaluate digital work while meeting the needs of this shifting landscape. This has begun to change, as leading scholars in digital history have taken up the clarion call for reform of the peer review process. The perception is that peer review has become outdated and stagnant, promoting conservative scholarship while also failing to exploit more dynamic means of communication and commentary through social media and Web 2.0 technologies. Robert B. Townsend, in a American Historical Association (AHA) blog posting proclaims, “The challenge lies in developing new forms of peer review better fitted to the online environment, both before publication (in the development and assessment stage) and after publication (as a means of validating the value and quality of the work).”[1] While Townsend and others have highlighted the ills of the current peer review system, they have yet to propose a set of review guidelines or methodology that would

bring together the interests and special knowledge of multiple disciplines.

In my DH2010 presentation, "Thinking Archivally: Selection, Search, and Reliability as a Framework for a New Digital Historiography," I proposed a framework for evaluating digital historical representations called *digital historiography*.[2] Digital historiography is the interdisciplinary assessment of digital historical representations across diverse formats. It promotes the rigorous and coherent use, recombination, and presentation of historical information through digital technologies. Digital historiography also accounts for the increased reliance on complex information systems to organize and represent historical data. The merging of historiography with archival theory and new media studies – along with numerous other fields such as museum studies, information science, and other humanities disciplines – ensures a comprehensive examination of a representation.[3]

Three fundamental areas of archival theory were raised in last year's presentation as fundamental to unlocking the trustworthiness and soundness of a representation: content selection, search functionality, and metadata. Through a series of illustrative examples of historical representations it was shown that these areas convey a representation's historical contextualization, where context is defined as the formal and humanistic relationships among data and resources. [4]

Whereas last year's paper focused on the theoretical underpinning of digital historiography, this year's presentation will provide a programmatic framework that scholars, archivists, librarians, curators, editors and technical specialists may adapt and apply to their own work. There are already a number of promising developments underway in establishing a set of guidelines and practices for digital humanities evaluation, particularly in literature; nonetheless, a similar approach in digital history has been lacking.[5]

This paper will provide a guide to evaluating digital historical representations using a series of exploratory questions. The following questions may provide an entry point with which to kick-start an analysis. It should be noted that these questions do not evoke clear-cut responses, but rather require deeper interpretation of the representation:

- Is the representation's historical content comprehensive or representative of the period/ event/issue in question?
- How do metadata schema and other descriptive information shape the interpretative possibilities of the representation?
- What capacity does the user have to repurpose historical data for additional study? To what extent does a user have sufficient contextual information such as the content's provenance to conduct such repurposing?
- How does the user search or navigate within the representation? In what ways does the interface facilitate or prohibit advanced humanistic inquiry?

These questions apply a combination of historical and archival understanding to address a representation's approach to selection, search, and metadata. At a more general level, they suggest the need within digital historiography to develop a peer review methodology that examines the *intersection* of technological and humanistic components of a representation.

2. A Three-Axis Framework for Peer Review

This paper will propose, based on the sample questions above, three interlocking axes for a peer review methodology: Historical Content, User Experience, and Creator Intent. Critiques of digital representations have all too often isolated one axis at the expense of considering the others. Scientifically driven user or human-computer interaction studies may overlook the unquantifiable nuances of historical contextualization. Similarly, an analysis of content often neglects a representation's user interface and how it may affect information access. Perhaps the most disregarded of the three areas is the creator's intent for constructing a digital historical representation. In the fierce competition for funding in the digital humanities, developers must be able to justify their resource-intensive project, which raises basic questions surrounding the representation's purpose and its contribution to historical scholarship or programming.

For the purpose of demonstrating how to apply digital historiography to peer review, a single digital historical representation will be selected and the audience will be guided through a brief, yet systematic evaluation. In terms of historical content, we will consider the representation's engagement with relevant scholarship. Although a digital representation's format may seem to belie traditional modes of historiography, it will be argued that a representation's demonstrated recognition of historiographic trends is essential for establishing whether it advances new areas of historical understanding. Unlike a text-based monograph, analysis of content depends on the

capacity of the user to *generate*, not just consume, plausible knowledge; therefore, peer review may focus more on the questions that may be posed through a representation rather than the construction of a single argument.

In terms of user experience, this presentation will focus on the question of access to historical content. More than just a scientific optimization of keywords, a reviewer must consider the user experience as a culmination of design features, tool functionality, and search capability. This raises issues regarding the application of standards and best practices. The selection of an appropriate metadata schema or digitization standard will determine a representation's interoperability with other resources, which will in turn influence historiographic assessment.

Finally, consideration of authorial intent, embedded within elements such as a site's introductory statement or contextual essays, provides a third axis that can anchor both a representation's content and user experience. The core question to be considered here will be the representation's intended or anticipated audience. Digital technology has enhanced access to historical knowledge for a wider public, which has placed additional responsibility on the part of the creator to consider how historical information is selected, presented, and most importantly, handled by diverse types of users.

3. The Benefits of Peer Review to Digital History

A peer review system that follows shared and accepted standards and methodology may have significant benefits for advancing digital historical scholarship and digital humanities infrastructure in general.[6] Representations may be assessed on the merits of content and usability. In terms of scholarship, digital historiography may liberate scholars and developers to consider how to harness technologies in service of historical inquiry when all too often the reverse seems to hold true. The time for experimentation for its own sake has passed, and we should begin to consider how technology might contribute to more sustainable development of innovative modes of scholarship.

This presentation is not intended to provide a fixed method for peer review; rather it will encourage a revival of humanities-driven interpretative analysis that addresses central areas of scholarship, audience, argument, and most importantly, inquiry. Peer review must be contingent on the unique set of questions related to a representation's subject area and formal qualities. Those in attendance will have the opportunity to take away a set of general questions that they may use to devise their own review criteria, or that may stimulate further dialogue about the peer review process. Perhaps most significantly, this presentation will defend the need for collaborative, transparent review that brings together subject and technical specialists, which can spark evaluation of noteworthy digital work that continues to elude mainstream academic recognition.[7]

Notes

1. Robert B. Townsend. "Assessing the Future of Peer Review." *AHA Today*. 7 June 2010. http://blog.historians.org/profession/1065/assessing-the-future-of-peer-review. The absence of peer review standards was also captured in a recent article for *The Chronicle of Higher Education* and online commentary. Jennifer Howard, "Hot Type: No Reviews of Digital Scholarship = No Respect," *The Chronicle of Higher Education* (2010), http://chronicle.com/article/Hot-Type-No-Reviews-of/65644/, accessed 31 October 2010. See also Dan Cohen. "Peer Review and the Most Influential Publications." 19 October 2010. http://www.dancohen.org/. For a more comprehensive discussion of digital peer review, albeit slightly outdated in its conclusions, see *Digital Scholarship in the Tenure, Promotion, and Review Process*, ed. Deborah Lines Andersen. Armonk: M.E. Sharpe, 2004.
2. The abstract for the DH2010 presentation may be found online in the conference program. *Digital Humanities 2010: Conference Abstracts*. Eds. The Alliance of Digital Humanities Organisations, The Association for Literary and Linguistic Computing, The Association for Computers and the Humanities, and The Society for Digital Humanities. London, 7-10 July 2010. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-747.pdf. Accessed 13 March 2011. An expanded version of the paper will appear in a forthcoming issue of *American Archivist*.
3. For a history and examination of the contemporary "divide" between the historical and archival professions, see Francis X. Blouin, Jr. and William Rosenberg. *Processing the Past: Contesting Authorities in History and the Archives*. Oxford University Press (2011).
4. For an introduction to the relationship of archival theory and context, see for example, Jennifer Meehan, "Towards an Archival Concept of Evidence," *Archivaria* 61(2006), 143. The discussion of archival context has been subsumed in a larger, more general area referred to as "information as evidence." For further discussion of this concept, see Terry Cook, "Archival Science and Postmodernism: New Formulations for Old Concepts," *Archival Science* 1(2001); Terry Cook, "What Is Past Is Prologue: A History of Archival Ideas since 1898, and the Future Paradigm Shift," *Archivaria* 43(1997); Margaret Hedstrom, "Archives, Memory, and Interfaces with the Past," *Archival Science* 2, no. 1-2 (2002); Joanna Sassoon, "Beyond Chip Monks and Paper Tigers: Towards a New Culture of Archival Format Specialists," *Archival Science* 7, no. 2 (2007); Jennifer Meehan, "Making the Leap from Parts to Whole: Evidence and Inference in Archival Arrangement and Description," *American Archivist* 72, no. 1 (2009).

5. Perhaps the most comprehensive guide to evaluating digital scholarship can be found with the MLA Guidelines for Editors of Scholarly Editions. Section V is devoted to electronic scholarly editions, and prompts reviewers to consider elements of a digital work such as TEI encoding and the user interface. An NEH-funded Summer Institute entitled "Evaluating Digital Scholarship" will be hosted by NINES at the University of Virginia 30 May – 3 June 2011, which will expand upon a one-day workshop held at Digital Humanities 2010 in London.
6. For further discussion of how the digital humanities must re-consider infrastructural needs, see Christine L. Borgman. "The Digital Future is Now: A Call to Action for the Humanities." Digital Humanities Quarterly. Fall 2009. Volume 3, Number 4
7. The thoughts and ideas expressed in this paper are entirely my own and do not reflect those of the National Endowment for the Humanities or any other federal agency.

You Suck at Narrative: Disciplinarity, Popular Culture, and the Database Logic of Photoshop

Stroupe, Craig
cstroupe@d.umn.edu
University of Minnesota Duluth

The series *You Suck at Photoshop* first appeared as a three-and-a-half-minute hoax circulating on the web in late 2007—a one-off Internet meme satirizing home-made, how-to videos ubiquitous on YouTube. Conceived and performed by advertising professionals Troy Hitch and Matt Bledsoe, the video featured an ironic, My-Last-Duchess-style commentary by a fictitious computer geek Donny Holye, who unintentionally reveals messy details of his marital, legal, and emotional life while demonstrating techniques of Photoshop. That same fall, in a special issue of *PMLA* on "Remapping Genre," some of the leading voices in the emergent field of the digital humanities discuss the database as a cultural form, and debate the political, ideological, and practical impact of databases on the work of the humanities, and the practices of narrative in particular. The varied contributions to "Remapping Genre" provide a dramatic tableau of twenty-first-century academic humanities confronting the database as a cultural discourse, and speculating on whether it represents an alien nemesis from the world of corporate capitalism, the long-sought embodiment of implicit literary ideals, or a value-neutral complement to narrative linearity.

This presentation argues that "Remapping Genre" and *You Suck at Photoshop* not only share an historical moment, but constitute a common critical effort to understand and respond to the historical rise of databases in cultural practice. Invoking this unlikely affiliation might seem to announce another scholarly incursion on the popular realm under the pirate flag of Cultural Studies. The relationship suggested here between academic and popular cultures, however, is less a territorial rivalry than a mutually transformational dialogue. And, in fact, this transformation is well underway. On one hand, as contributors to "Remapping Genre" amply document, literary studies is rapidly adopting frictionless, digital tools to perform its daily work: from personal, bibliographic browser plug-ins like Zotero, to online databases like The *Walt*

Whitman Archive, to vast digitalization initiatives like Google Books. Less recognized, on the other hand, is the degree to which popular, database-driven New Media like Facebook, online games, and the web itself are increasingly being constructed not only by the work of programmers and designers, but by narratives that users, players, and an emerging circle of professional critics employ to shape, remember, and publicize their experiences of virtual database environments. Such narratives reveal how the texture of language—the friction of fiction—is an integral part of how databases are culturally constructed and experienced, and the extent to which that use of narrative language is conditioned by a literary sensibility.

This talk will analyze clips from *You Suck at Photoshop*—as well as background on the series' circulation and presentation on the web—to examine the interplay of narrative and database as textual and cultural logics, as well as to suggest the historical transformations that such an emergent dialogic is producing in the relationship between the academic humanities and popular culture. Like Donny Hoyle's conflicted relationship with Photoshop's remorselessly vast database of tools, the Digital Humanists featured in "Remapping Genre" express ambivalent responses to database-driven tools of their own academic trade, and debate Lev Manovich's claim that database and narrative are "natural enemies…[c]ompeting for the same territory of human culture, each claim[ing] an exclusive right to make meaning out of the world" (225). Ed Folsom's and Kenneth Price's digital project The *Walt Whitman Archive* provides a common case in point, serving to ground, historicize, and set in relief the critical differences among these scholars, which echo concerns and anxieties within the profession at large: questions concerning the degree to which databases may or may not threaten narrative forms of thought and meaning (N. Katherine Hayles); the ideological, as opposed to the supposedly libratory, effects of databases, especially mediated by large, corporate interfaces like Google (Jonathan Freedman); the extent to which the seemingly new logic of databases remediates pre-digital, even ancient literacies (Peter Stallybrass and Meredith L. McGill); whether or not the database can be called a genre, perhaps even belonging to the "epic" genre that crosses national and historical boundaries (Wai Chee Dimock); and the problems of deciding whether a tool like The *Walt Whitman Archive* is really a database at all (Jerome McGann).

Beyond particular insights into the database/ narrative question, this talk will explore how the very unlikeliness of this dialogue of You Suck at Photoshop and "Remapping Genre" points to emergent affinities and even alliances among popular, academic, and economically "productive" discourses. If these potential alliances have not been sufficiently described, it is not from lack of trying, however. In his monumental book *The Laws of Cool*, for instance, literary critic and digital theorist Alan Liu attempts to imagine the relationship between the cultures of "the literary"— not just "works of literature as such," he says, but "the underlying sense of the literary"—and of digital "knowledge work" characterized by an ethos of "cool" (1, 3). "In [capitalism's] regime of systematic innovation [and creative destruction]," he asks at the beginning of the book, "is the very notion of the literary doomed to extinction if —or, rather, especially if—it dares to imagine a literature of the database, spreadsheet, report, and Web page?" (3). Essentially, Liu is questioning just how "the contemporary humanities and arts…[might] not only make contact with the generations of cool but lead them beyond the present limitations of cool" (381).

In his somber Epilogue, Liu declares, "I am a believer at heart…. I think literature will indeed have a place in a new-media world…. But what the eventual nature and position of that literature will be among the convergent data streams of the future is something I do not yet know how to theorize" (389). Like Richard Lanham, Katherine Hayles, and other scholars leading the search for the Holy Grail of would-be database literature, Liu tends to look for examples in high-cultural practices of Internet Art, or what he more inclusively terms "ethical hacking," but Liu ultimately finds them "too closely associated with anarchist, Situationalist, radical leftist, and/or high-theoretical paradigms…to offer persuasive models for an art that might affect the knowledge worker in his or her ordinary cubicle" (397-98). This talk will argue that such persuasive models, or perhaps *models* of models, can be found in the ordinary cubicles of slackers like Donny Hoyle—but only if academic culture can develop the critical idiom to describe them.

References

Folson, Ed (2007). 'Database as Genre: The Epic Transformation of Archives.'. *PMLA.* 122: 1571-1579.

Liu, Alan (2004). *The Laws of Cool: Knowledge Work and the Culture of Information. 1st ed..* University Of Chicago Press.

Manovich, Lev (2002). *The Language of New Media. 1st ed..* Cambridge Mass: MIT Press.

Medical Case Studies on Renaissance Melancholy: Online Publication Project

Suciu, Radu
risuciu@gmail.com
Université Paris-Diderot France

This paper presents the intermediary results of our ongoing research project at the Université Paris-Diderot as part of a post-doctoral bursary awarded by the Mayor of Paris's Research in Paris 2010 program. The project aims at combining traditional research methods (research, annotation and publication of early modern texts and documents) with open source tools and standards (Omeka, Zotero, TEI), with the goal of publishing an online encyclopedia of case studies on medicine and melancholy in the late Renaissance.

2. Historical Background

The principal research question asked is: how did the Renaissance physician position himself in relation to his patient, and how does he attempt to document his 'clinical' experiences in writing? The case histories of those suffering from melancholy are instrumental in understanding this issue: tormented by various hallucinations and deliria, the melancholy see what is not there and live in a world of strange delusion, variously believing that they have no head, or are made of brick, or of butter, and so forth. The patient who famously believed his body to have been transformed into butter feared even approaching the oven (an awkward situation since his line of work was in baking bread), while yet another was convinced he was missing one leg, bitten off by an imaginary crocodile. Cases such as these are at the heart of our research: we have examined not only early modern medical documents, but also many important collections of commonplace books in our search for case studies, patient descriptions, medical observations, and so-called 'curative epistles'.

3. Online Publication of the Research Materials

Rather than a traditional publication in print, the results are being progressively published online with the aid of a number of open source tools. The principal aims of this paper are to present the various preparatory research stages, the choices made in implementing the digital methods and tools, and finally to reflect on the evolution of the project in the years to come.

4. Methods and Tools: TEI Transcription, Omeka, Zotero Integration

This project uses the TEI recommendations for the transcription and the encoding of early modern medical texts. The TEI has been demonstrated to be the most comprehensive way of transcribing rich, complex texts by a number of major projects [1]. TEI documents are then stored in an online database that uses and adapts the open source CMS Omeka [2], now a standard tool for the creation of online repositories of documents and virtual exhibitions. We shall present the way in which we have used and adapted Omeka's plugins and themes. We shall also discuss the metadata structuring choices we have made. Since this is handled directly by Omeka, it facilitates the creation of an OAI repository, which can be made directly accessible to data-harvesters [3] and eases integration with research management applications such as Zotero.

5. Conclusion: Putting the University Database on Virtual Exhibition

The textual documents transcribed and added via the Omeka database are to be accompanied by critical annotations, literary transpositions or references and by a collection of images or an index of commonplaces. The website, unlike a scholarly publication, will be more easily accessible and reach a wider audience, while the database, making use of Web 2.0 technologies, will function as a virtual exhibition, an online 'cabinet of curiosities', allowing readers to interact with, comment on and contribute to the published materials. With this in mind, the ultimate aim of this digital humanities project is to generate a broader interest in early modern research and history by focusing on melancholy, a subject that has never ceased to influence and inform disciplines from medicine to literature.

References

Blair, Ann (2010). *Too Much to Know: Managing Scholarly Information before the Modern Age.* New Haven: Yale University Press.

Bugei, Nyaosi (2010). 'Voltaire's Correspondences. Utilizing Visualization in the Mapping the Republic of Letters Project'. *Spatial History Lab.* September

2010. http://www.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=71.

Burnard, Lou, Sperberg-McQueen, M. (2006). *TEI Lite: Encoding for Interchange: an introduction to the TEI.* http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html.

Cohen, Daniel, Rosenzweig, Roy (2005). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web.* Philadelphia: University of Pennsylvania Press. http://chnm.gmu.edu/digitalhistory/index.php.

Dacos, Marin (ed.) (2010). *Read-write book: le livre inscriptible.* Marseille: Cléo.

Dandrey, Patrick (2005). *Anthologie de l'humeur noire. Écrits sur la mélancolie d'Hippocrate à l'Encyclopédie .* Paris: Le Promeneur.

Ferrand, Jacques (1990). *A Treatise on Lovesickness (1623).* Beecher, Donald A., Ciavolella, Massimo (eds.). Syracuse: Syracuse University Press.

Findlen, Paula (1994). *Possessing Nature: Museums, Collecting, and Scientific Culture in Early Modern Italy.* Berkeley: University of California Press.

Grafton, Anthony (2008). *Codex in Crisis.* New York: Crumpled Press.

Kucsma, Jason, Reiss, Kevin, Sidman, Angela (2010). 'Using Omeka to Build Digital Collections: The METRO Case Study'. *D-Lib Magazine.* 3/4. http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/march10/kucsma/03kucsma.html.

Schreibman, Susan, Siemens, Ray, Unsworth, John (eds.) (2004). *A Companion to Digital Humanities.* Oxford: Blackwell. http://www.digitalhumanities.org/companion/ (accessed March 10, 2011).

Notes

1. See for example the ongoing Transcribe Bentham: A Participatory Initiative. Available at: http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham [Accessed October 4, 2010].
2. Omeka is a project of the Center for History and New Media, George Mason University. Available at: http://omeka.org/about/ [Accessed March 10, 2011].
3. See the forthcoming 'Isidore' developed by the French Centre National de la Recherche Scientifique (CNRS): ISIDORE - Accès aux données et services numériques de SHS. Available at: http://www.rechercheisidore.fr/ [Accessed March 10, 2011].

A User-Centered Digital Edition of Vuk Stefanović Karadžić's *Lexicon Serbico-Germanico-Latinum*

Tasovac, Toma
ttasovac@humanistika.org
Center for Digtial Humanities (Belgrade, Serbia)

Ermolaev, Natalia
ne99@columbia.edu
Center for Digtial Humanities (Belgrade, Serbia)

Dictionaries lie at the core of the human ability to conceptualize, systematize and convey meaning. But a dictionary (both print and digital) is many things at once: a text, a tool, a model of language, and a cultural object deeply embedded in the historical moment of its production (Tasovac, 2010). While it is true that we now live in the age of the electronic dictionary (de Schryver, 2003), dictionaries have always played an important role in the interplay between production technology and knowledge taxonomies (McArthur, 1986; Hüllen and Schulze, 1988; Hüllen, 1999). In this respect, historical dictionaries remain particularly valuable objects of study because they illustrate sociolinguistic perceptions and reveal culturally shaded conceptualizations of lexical knowledge of a particular epoch — often in stark contrast to our contemporary attitudes and values. Moreover, they pose a veritable challange for text encoding, semantic markup and database modeling (Fomin and Toner, 2006; Nyhan, 2006; Nyhan, 2008; Mooijaart and van der Wal, 2009; Lemnitzer et al., 2010). This is why all dictionaries, including retrodigitized historical dictionaries, are important for digital humanities, and why DH — with its concern for (abstract) modeling of knowledge and its (practical) implementations in humanities research — can integrate and propel different trains of lexicographic and metalexicographic thought at the intersection of language and technology.

Many DH research projects have aimed to produce electronic editions of printed lexicons (see for instance Morrissey, 1993; Lemberg et al., 1998; Christmann, 2001; Fournier, 2001). In such efforts, retrodigitization is usually based on one of two approaches: either the production of "faithful" digital copies (at the cost of reproducing factual or typographic errors), or the structural modelling of the content, which

treats the print edition as a data source, rather than as an immutable text to be reproduced in its entirety (Lobenstein-Reichmann, 2008). In either case, retrodigitization projects tend not to involve any degree of re-editing or expanding the actual content of historical dictionaries.

We agree with Kirkness (2008) that digitalizing historical dictionaries can increase and optimize their use value, especially in global, networked environments. But we also feel that one central aspect is often overlooked in current studies of retrodigitized dictionaries: users interacting with a historic lexicon do no longer necessarily have active command of the text's primary language. Even when historical dictionaries are retrodigitized with the user's needs in mind, the focus is usually on easy-to-handle navigation, presentation layout and retrieval of elements from a full-text search (Christmann, 2003); or on uniformization of existing data elements, such as dates (Kinable, 2006). While these efforts are worthwhile and necessary because they contribute greatly to editions that are more usable and efficient than their hardcopy counterparts, electronic dictionaries remain in essence lookup tools (for words encountered in a given text) rather than exploratory tools (for unknown words or concepts). This, we believe, can reduce both their scholarly and popular appeal.

It may seem unlikely at first that historical dictionaries can generate non-academic interest, but experience has shown that there is a broad audience outside highly professionalized linguistic circles that is both curious and enthusiastic about exploring the historic and ethnographic fabric of a language (Kirkness, 2008). In our own web-project — “Reklakaza.la” (Serbian for “hearsay”) — we have been publishing online selected entries from the classic, 19th-century Serbian Dictionary by Vuk Stefanović Karadžić and linking them via social networks Twitter (http://twitter.com/ Vuk_Karadzic) and Facebook (http://facebook.com/reklakaza.la). The project has gained more than 24,000 fans on Facebook alone, becoming a platform for bringing meaningful humanities inquiry into the public conversation, fostering the sense of community, sharing, and mutual learning that proves the relevance of the humanities in today's world despite academic budget cuts and declining job opportunities.

The success of our pilot project has strengthened our conviction that a modern, electronic edition of Karadžić's dictionary is long overdue. Vuk Stefanović Karadžić (1787-1864), the linguist, folklorist and reformer of the Serbian language, published his landmark *Srpski rječnik, istumačen njemačkijem i latinskijem riječima — Lexicon Serbico-Germanico-Latinum* in two editions (1818 and 1852). This first lexicon of the modern Serbian vernacular, rather than the Church-Slavic hybrid language used by the educated elites up to the 19th century, has a unique place in the history of not only the Serbian language, but the South Slavic diasystem in general (Дмитриев and Сафронов, 1984; Wilson, 1986; Стојановић, 1987; Eschker, 1988; Potthoff, 1990; Ивић, 1990; Vitalich, 2005; Кулаковский, 2005). The text is rich with ethnographic and anthropological material. Not only do many entries contain examples of Balkan folk storytelling, but some are themselves structured as historical, cultural and ethnographic narratives that offer informative sketches and sometimes even very detailed accounts of the myths and realities of the Balkan past (see, for instance, entries for кмет, отмица, мора, хајдук, etc.).

Though it was republished twice (in 1898 and 1935), the *Lexicon* has not been reprinted since (other than in facsimile editions). Meeting the needs of modern-day users, however, presents a host of editorial challenges. The entries in the *Lexicon* are written mainly in a dialect which is on the margins of contemporary standard Serbian. Thus, the lexicographic material is not always entirely understood by contemporary speakers, and can often appear obscure or unwieldy. It is hard for the average user to answer questions such as: what was the early 19th-century equivalent of a mod-ern-day Serbian word? What household objects, for instance, are listed in Karadžić‘s dictionary? What words were difficult or impossible for Karadžić to translate into German or Latin?

Our “Annotated Digital Edition of Vuk Stefanović Karadžić’s *Srpski rječnik*” is therefore conceived as a resource that caters to access needs and habits of modern scholars, teachers, students, and, last but not least, general readers. The entries are marked up XML, in compliance with the Guidelines of the Text Encoding Initiative (Burnard et al., 2006). In this, initial phase of the project, we are focusing solely on text encoding, but in view of the potential use in a data-base driven web-application at a later stage.

In addition to marking up existing structural elements of a dictionary entry (such as lemma, part of speech, senses, definitions, translation equivalents, examples etc.), our work supplies important additional information that will enhance the modern-day user's interaction with the dictionary, including:

- standard Serbian equivalents to dialect word forms (e.g. бичевање vs. бичкарење, мешина vs. мљешина, енглески vs. инглешки);
- Serbian ekavian word-forms to both standard and east-Herzegovina jekavian entries (e.g. терати vs. тјерати, терати vs. ћерати);
- both the original 19th-century accentuation (e.g. кȍчијāшки) and its modern-day graphic equivalent (кȍчијāшкӣ);
- indications when modern-day accentuation differs from the from found in the *Lexicon* (e.g. мо̑ре vs. мо̀ре, das Meer, mare);
- an extension of the extant cross-reference system through linking synonymous and near-synonymous entries that have been overlooked by previous editors (e.g. жаба and напнигуша; обрљуга and неопера).
- labeling of Turkisms overlooked by previous editors (e.g. була, инћар, џукела);
- marking up persons, places and dates for easy indexing and analysis;
- indications of word usage (eg. ист. and ист. кр. as `<usage type="geo">East</usg>` for better statistical analysis and possible further processing and creation of geo-spatial word maps etc.);
- marking up instances where Karadžić uses a first-person narrative to explain an entry;
- indications of the edition in which entries appeared for the first time, etc.

Furthermore, we are assigning semantic domain labels to word senses in accordance with Magnini and Cavaglia, 2000; Bentivogli et al., 2004, cross-referencing senses with the Transpoetika Dictionary — a bilingualized, Wordnet-based Serbian-English dictionary (Tasovac, 2009), and providing English glosses in addition to the existing German and Latin. All of this will help us meet our goal of moving beyond the current paradigm of limiting retrodigital text editing to the creation of electronic replicas of hardcopy lexicons or semantically structured electronic representations of the original data source. We are interested in hybrid approaches that respect the integrity of the original text, but also take advantage of the digital medium to create modern, deeply-encoded, user-centered editions of historical dictionaries, which can not only provide look-up mechanisms for particular words, but also function as exploratory tools for various types of knowledge discovery.

Some practical advantages of our edition of Vuk Karadžić's dictionary will include reverse look-ups, allowing a user to search an English, German, or Latin word and find its Serbian equivalent in the Lexicon. The domain labels will provide researchers with valuable and, for the first time, measurable information about the clusters of paradigmatically related terms, as well as the extent of domain ambiguity and domain variability. Users will be able to treat semantic domains as thematic entry points into the dictionary, looking up, for instance, all entries that belong to **AGRICULTURE**, **FOLKLORE**, **HISTORY** or **GASTRONOMY**; while our logically and semantically consistent markup of Karadžić's own usage notes will make it possible for users to easily explore regional and dialectological distribution of entries in the *Lexicon*, offering a basis for subsequent work that could involve data visualization, statistical analysis, text mining etc.

References

Bentivogli, L, P Forner, B Magnini, E Pianta (2004). 'Revising the Wordnet Domains Hierarchy: Semantics, Coverage and Balancing'. *Proceedings of the Workshop on Multilingual Linguistic Resources.* Pp. 101-108.

Burnard, Lou, Katherine O'Brien, John Unsworth (2006). *Electronic Textual Editing.* New York: The Modern Language Association of America.

Christmann, R (2001). 'Books into bytes: Jacob and Wilhelm Grimm's Deutsches Worterbuch on CDROM and on the Internet'. *Literary and Linguistic Computing.* no. 2: 121-133.

Christmann, R (2003). 'Towards the User: The Digital Edition of the Deutsche Worterbuch by Jacob and Wilhelm Grimm'. *Literary and linguistic computing.* 1: 11-221 : . .

de Schryver, Gilles-Maurice (2003). 'Lexicographer's Dreams in the Electronic-Dictionary Age'. *International Journal of Lexicography.* 2: 143-1992 : .

Eschker, Wolfgang, ed. (1988). *Jacob Grimm und Vuk Karadžić: Zeugnisse einer Gelehrtenfreundschaft.* Volkskundliche Schriften. Kassel: E. Röth-Verlag V. Bd. 4. .

Fomin, Maxim, Gregory Toner (2006). 'Digitizing a Dictionary of Medieval Irish: the eDIL Project'. *Literary and Linguistic Computing.* 1: 83.

Fournier, Johannes (2001). 'New directions in Middle High German lexicography: dictionaries interlinked

electronically'. *Literary and Linguistic Computing.* no. 1: 99-111. .

Hüllen, Werner (1999). *English Dictionaries, 800-1700: The Topical Tradition.* Oxford [England] New York: Clarendon Press Oxford University Press.

Hüllen, Werner, Rainer Schulze (1988). *Understanding the Lexicon: Meaning, Sense, and World Knowledge In Lexical Semantics.* Tübingen: M. Niemeyer.

Kinable, Dirk (2006). 'Computerized Restoration of Historical Dictionaries: Uniformization and Date-assigning in Dictionary Quotations of the Woordenboek der Nederlandsche Taal'. *Literary and Linguistic Computing.* . .

Kirkness, Alan (2008). 'Digitalisierung -Vernetzung -Europäisierung: Zur Zukunft der historischen Lexikographie des Deutschen'. *Lexicographica.* 7-38.

Lemberg, I, S Petzold, H Speer (1998). 'Der Weg des Deutschen Rechtswörterbuchs in das Internet'. *Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium,.* Tübingen: Niemeyer, pp. 262-284.

Lemnitzer, Lothar, Laurent Romary, Andreas Witt (2010). 'Representing Human and Machine Dictionaries in Markup languages (SGML, XML)'. `http://arxiv.org/pdf/0912.2881`.

Lobenstein-Reichmann, Anja (2008). 'Allgemeine Überlegungen zur Retrodigitalisierung historischer Wörterbücher des Deutschen'. *Lexicographica.* 173-198.

Magnini, B, G Cavaglia (2000). 'Integrating Subject Field Codes into WordNet'. *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation.* Pp. 1413-1418.

McArthur, T. B. (1986). 'Thematic Lexicography'. *The History of Lexicography: Papers from the Dictionary Research Centre Seminar at Exeter, March 1986.* R. R. K Hartmann (ed.). Amsterdam; Philadelphia: J. Benjamins, pp. 40.

Mooijaart, Marijke, Marijke van der Wal (eds.) (2009). *Yesterday's Words: Contemporary, Current and Future Lexicography.* Historiographia Linguistica. Newcastle: Cambridge Scholars Publishing.

Morrissey, Robert (1993). 'Texts and Contexts: The ARTFL Database in French Studies'. *Profession.* 27-33.

Nyhan, Julianne (2006). 'The Application of XML to the historical lexicography of Old, Middle, and Early-Modern Irish: a Lexicon based analysis'. Cork: National University of Ireland. .

Nyhan, Julianne (2008). 'Developing Integrated Editions of Minority Language Dictionaries: The Irish Example'. *Literary and Linguistic Computing.* 1: 3-121 : .

Potthoff, Wilfried (ed.) (1990). *Vuk Karadžić im europäischen Kontext.* Beiträge des internationalen wissenschaftlichen Symposiums der Vuk-Karadzic-Jacob-Grimm-Gesellschaft am 19. und 20. November 1987. Heidelberg: Carl Winter Universitätsverlag.

Tasovac, Toma (2010). 'Reimagining the Dictionary, or Why Lexicography Needs Digital Humanities'. *Digital Humanities 2010.* `http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-883.html/`.

Tasovac, Toma (2009). *More or Less Than a Dictionary? Wordnet as a Model for Serbian L2 Lexicography. Infotheca: Journal of Informatics and Librarinaship.* no. 1-2: 13a-22a.

Vitalich, Kristin Leigh (2005). 'Lexicographical doxa: The writing of Slavic dictionaries in the nineteenth century (Samuel Bogumil Linde, Vuk Stefanović Karadžić, Vladimir Ivanovich Dal)'. University of California at Los Angeles.

Wilson, D (1986). *The life and times of Vuk Stefanović Karadžić, 1787-1864: Literacy, Literature, and National Independence in Serbia.* Oxford: Clarendon Press.

Дмитриев, Петр Андреевич, Герман Иванович Сафронов (1984). *Вук С. Караджич и его реформа сербскохорватского/хорватосербского литературного языка. Учеб. пособие.* Ленинград: ЛГУ.

Ивић, Милка (1990). *О језику Вуковом и вуковском.* Нови СадКњиж. заједница Новог Сада.

Кулаковский, Платон Андреевич (2005). *Вук Караджич. Его деятельность и значение в сербской литературе.* Москва: УРСС.

Стојановић, Љубомир (1987). *Живот и рад Вука Стеф. Караџића.* Београд: Београдски издавачко-графички завод.

Probabilistic Analysis of Middle English Orthography: the Auchinleck Manuscript

Thaisen, Jacob
jthaisen@gmail.com
University of Stavanger

The bulk of the literary materials that survive from the Middle English period are scribal copies, rather than authorial compositions. Such copies pose a challenge to the stylometrist, the reason being that copies written in a single scribal hand may have non-identical orthographic profiles. Their non-identity is a product of their transmission history, as the typical copy will contain both spelling forms originating in the exemplars and other such forms introduced by the scribe. Historical dialectologists have developed methods for separating the mix of scribal usage and exemplar usage typically recorded in a single scribal copy. Although powerful, these methods rely on questionnaire-based interrogation of text samples and subsequent visual analysis of spelling forms arranged in tables. The arrival of digital transcripts has sped up the data collection process and has led to compilation of fuller profiles, but the questionnaire itself has stayed. Thus, these methods fail to take full advantage of the digital medium.

This presentation demonstrates that a purpose of identifying and isolating locations in which the makeup of the spelling system changes during the full text of a longer Middle English literary manuscript may be met by probability-based comparison of text samples. What spelling forms happen to be attested in a given text is a function of what words happen to make up that text. The direct comparison of texts is therefore not readily possible. It has typically been made possible by considering profiles recording only the spelling forms of those words which may reasonably be expected to occur in every text, such as function words like "such", "that", and "these". The alternative solution proposed in this paper is to base assessment of similarity on "models" of text samples' spelling–exhaustive profiles of which letters and letter sequences occur in them and with which frequency. I shall refer to single letters as unigrams, ordered sequences of two letters as bigrams, etc. Such models are easily compiled from electronic diplomatic transcripts. The dissolution into n-grams is equal to identification of the between texts because comparison of the building blocks is in itself relatively independent of the word level.

Similarity is measured between a text sample (the test sample) and a model derived from another text sample (the training sample). It is expressed as the overall probability that the test sample is an instance of the same spelling system as the system modelled. Computing this probability proceeds, with a trigram model, from consideration of each unigram in the context of the bigram preceding it–the reader may correctly recognise the "Markov assumption" in this description. What is output, however, is the reciprocal of the average probability per gram. This entity, called "perplexity", will conveniently always be a positive number larger than 1 with the present type of data. Moreover, techniques exist for "smoothing" a model, that is for reducing its dependence on the words constituting the training sample. This reduction is achieved by statistically manipulating the probabilities computed for the training sample n-grams. Smoothing additionally leads to probability being assigned to spelling forms unattested in the training sample.

It is these properties of these techniques that makes their application on n-gram models based on Middle English texts further increase the comparability of those texts. Probabilistic modeling techniques have, however, as far as I am aware, rarely been applied for the stylometric analysis of Middle English materials, and it has yet to be established which specific smoothing technique produces the most satisfactory models of those materials.

A simple example may illustrate. The spelling forms <suich, suyche, such> for the word "such" are found in text A, while the same word is spelt <suche> in text B. Intuitively, the text B spelling form <suche> falls within the range of variation characteristic of the spelling system of text A but happens to be unattested in it, while other known Middle English forms such as as <swylke, suilk> do not. The present methodology involves dissolving <suche> into <su, suc, uch, che, he> and establishing the smoothed conditional probability for each of these trigram building blocks in text A (thus obtaining a intuition that the spelling form is possible in the spelling system of text A. In practice, however, the quantification is not effected for the individual form but for the whole of text B in relation to text A.

To illustrate the adequacy of perplexity-based comparison in stylometry, I trace changes in spelling in a large manuscript collecting several Middle English literary works. The corpus is the Auchinleck manuscript, Edinburgh, National Library of Scotland,

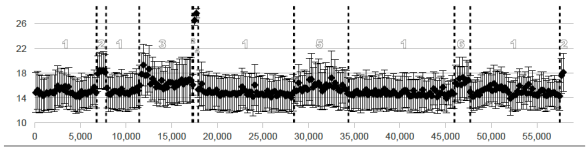
Advocates' MS 19.2.1, produced in the London-Westminster area in the first half of the fourteenth century. The potential influences on the scribes include the literary structure, as the codex's total of almost 59,000 lines of text are divided between no less than forty-four literary works representing a range of genres. A map showing locations in which the spelling system changes during the full text of the Auchinleck manuscript may be expected to reflect the literary structure only if the exemplars did so and the Auchinleck scribes reproduced them slavishly. By contrast, it is the boundaries of the scribal contributions that will be visible in the map if each scribe thoroughly converted into his own spelling system when copying. Six scribal hands are present. Of these, Scribe 4's contribution is too short (551 words) to constitute a reliable sample, while the usages of the other five scribes should be visible. They are distinct in terms of their typological classification on the dialect continuum, although they fall into an eastern cluster and a western one. Dialect analysis has thus placed Scribe 1 in Middlesex, Scribe 2 on the Gloucestershire-Worcestershire border, Scribe 3 in London, Scribe 5 in Essex, and Scribe 6 in Worcestershire (McIntosh, Samuels, et al 1986, I: 88; LPs 6510, 6940, 6500, 6350, and 7820).

To be able to compute and compare probability for sections of the Auchinleck manuscript against one another, I obtained a digital transcript of its text from the Oxford Text Archive (Burnley and Wiggins 2003). The transcript is suitable, because rather than modernise the spelling forms found in its source, it reproduces the source in conformity with standard practice of diplomatic transcription. My tool for constructing models and computing perplexity is the SRI Language Modeling Toolkit (SRILM; Stolcke 2002); this toolkit is freely available for noncommercial purposes from the website of its SRILM constructed and smoothed an interpolated model for every 200-line section; the smoothing technique selected was that described by Witten and Bell (1991). This technique was developed for purposes of text compression at the level of the word but it is appropriate for application on Middle English spelling data too. The reason is that the technique effectively assigns probability to collocating letters as if they were a single letter, rather than a series of independent units.

The toolkit took the same modified transcripts of all the sections as the input and computed their similarity with the models. The computation resulted in a separate model for each section, and for every such model, a separate perplexity for every section. I established the mean perplexity and standard deviation for the perplexities obtained for each model. The box and whisker graph below shows the results. In this graph the vertical axis gives perplexity and the horizontal axis position in the text of the Auchinleck manuscript. The diamond represents mean perplexity and the T-bar represents half a standard deviation, so that one upright T-bar and its reverse together indicate an interval of one standard deviation from the mean. A dashed vertical line appears for ease of reference at the boundary of a scribal stint as established by palaeographers (Bliss 1951), with the outlined numbers identifying the scribes.

As is apparent, the figure distinguishes the scribes of the Auchinleck manuscript. The rises and falls in mean perplexity during the text strongly correlate with the boundaries of the scribal stints, while mean perplexity is relatively constant within every such stint. Repetion of the experiment with other divisions of the text produced results sufficiently similar to Figure 1 to establish the pattern as being a property of the data rather than an artefact of the method.

It would have been time-consuming indeed to conduct a questionnaire-based interrogation of the full text of the Auchinleck manuscript. Visual analysis of the resulting profile to identify and isolate locations in which the spelling system changes would, moreover, have been complex, because of the amount of data and difficulty of isolating the diagnostic features. Perplexity-based comparison as illustrated above, by contrast, requires little preprocessing of the transcript, is effected in an afternoon, and is based on all the available data.



Mean perplexity and standard deviation in the Auchinleck manuscript (trigram models, 200-line sections)

References

Bliss, A. J. (1951). 'Notes on the Auchinleck Manuscript'. *Speculum.* 26: 652–58.

Burnley, D., A. Wiggins (eds.) (2003). *The Auchinleck Manuscript.* `http://www.nls.uk/auchinleck`.

McIntosh, A., M. L. Samuels, M. Benskin (eds.) (1986). *A Linguistic Atlas of Late Mediaeval English.* Aberdeen: Aberdeen University Press V. 4 vols. .

Stolcke, A. (2002). *SRILM: An extensible language modeling toolkit. Proceedings of the 7th International Conference on Spoken Language Processing.* Hansen, P., Pellom, B. (eds.). Denver: Casual Productions.

Witten, I. H., T. C. Bell (1991). *The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. IEEE Transactions on Information Theory.* , pp. 1085–94 37: 1085–94.

Opening the Gates: A New Model for Edition Production in a Time of Collaboration

Timney, Meagan
mbtimney.etcl@gmail.com
Electronic Textual Cultures Laboratory University of Victoria

Leitch, Cara
cmleitch@gmail.com
Electronic Textual Cultures Laboratory University of Victoria

Siemens, Ray
siemens@uvic.ca
Electronic Textual Cultures Laboratory University of Victoria

In the very early days of the world wide web, but well into a period in which our community understand the positive and transformative impact that computational technique has had on scholarly editing, Fortier reminded us that literary studies is and always has been focused on the study of texts regardless of interpretive theoretical predisposition. In digital literary studies, that textual focus manifests in a number of theories about the nature of the text in general and the electronic scholarly edition in particular, and has developed such that a basic typology of electronic scholarly editions is relatively straightforward to construct via the approach taken in handling textual materials. Well into what is often called the age of Web 2.0, it is worth noting that prominent types of electronic scholarly editions were largely developed before the ubiquity of the world wide web that we now enjoy and do not accurately reflect its current academic engagement. Indeed, given that we have now entered a new phase in the *social* formation of the web, we can no longer ignore the influence of new networks and connections on the scholarly digital edition. Our understanding of the electronic scholarly edition requires reconsideration in light of the collaborative potential of current and emerging digital technologies; put another way, we need to extend our typology in light of new models of edition production that embrace social networking and its commensurate toolkit. We propose that, while the digital medium is most certainly a productive space in which to analyse editions (as proposed by Hans Walter Gabler), the social incarnation of the digital edition allows us to

refocus our systematic analysis of *texts*, thus furthering the reconfiguration of the hierarchy for reading both texts and editions.

This working paper offers a new understanding of the historical underpinnings of the scholarly and digital editions, and envisions the possibilities of the scholarly social edition. It is generally accepted that there are several basic models for electronic editions of a scholarly nature, each put forward before the advent of the world wide web -- each demonstrating disparity within and among approaches in handling the text that lies at their centre. Using Unsworth's scholarly primitives as a model for describing the set of activities common to humanities scholars, we have developed a functional definition for the strategies employed by expert readers: Analysis, Synthesis, Communication, and Dissemination. New methods of engagement are both social in nature and reflect the interrelated nature of these strategies: analysis and synthesis grow from communication that, in turn, affects dissemination, and so forth. Based on recent research concerning the reading strategies of expert or professional readers, and the current state of digital humanities scholarship, the next step in the development of the scholarly edition is one that reflects the importance of collaboration, incorporates contributions by its readers, and where the editor acts as a facilitator for user involvement rather than enjoying an unassailable final word. Our model of the social edition points to new methods of engagement in digital literary studies. The social edition embraces the collective (but without losing sight of the individual) and accepts that no edition is ever truly complete.

Despite Stephen Nichols's call to 'dismantle the silo model of digital scholarship,' many digital editions, like print editions, continue to exist as self-contained units that do not encourage interaction with other resources. Instead we would argue that the social edition grows from Greg Crane's exhortation: '[w]e need to shift from lone editorials and monumental editions to editors as … editors, who coordinate contributions from many sources and oversee living editions.' The movement toward social edition production has already begun, with projects such as EEBO interactions, 'a social networking resource for *Early English Books Online*' (`http://eebo-interactions.chadwyck.com`) and George Mason University's 'Crowdsourcing Documentary Transcription: an Open Source Tool,' (`http://scripto.org/`) which is described as 'an open source tool that would allow scholars to contribute document transcriptions and research notes to digital archival projects, using the Papers of the War Department as a test case.' These projects, among others, point to a growing need in the scholarly community to expand our knowledge communities using the social technologies at our disposal. With the understanding that we cannot prophesize the exact nature of the social edition at this current juncture, we do, however, wish to reiterate the importance of seeing the scholarly text as a process, and the initial, primary editor as a facilitator, rather than progenitor, of knowledge creation.

References

Avram, G. (2006). 'At the Crossroads of Knowledge Management and Social Software'. *The Electronic Journal of Knowledge Management.* 4.1: pp. 1-10.

Bouman, Wim et al. (2008). 'The Realm of Sociality: Notes on the Design of Social Software'. *Sprouts.* `http://sprouts.aisnet.org/8-1/` (accessed 30 April 2010).

Boyd, Stowe. (2006). 'Are You Ready for Social Software?'. */message.* (accessed 30 April 2010).

Bryant, Todd. (2006). 'Social Software in Academia'. *Educause Quarterly.* 61-64.

Crane, Greg. (2010). 'Give us editors! Re-inventing the edition and re-thinking the humanities'. 'The Shape of Things to Come'. Charlottesville, VA. `http://shapeofthings.org/papers/` (accessed 30 April 2010).

Elia, Gianluca, Angelo Corallo (2009). 'A Knowledge Strategy Oriented Framework for Classifying Knowledge Management Tools'. *Knowledge Networks: The Social Software Perspective.* Miltiadis D. , Lytras, Robert Tennyson, Patricia Ordonez de Pablos (eds.). Hershey, PA: Information Science Reference, pp. 1-16.

Faulhaber, Charles B. (1991). 'Textual Criticism in the 21st Century'. *Romance Philology.* 123-48.

Fitzpatrick, Kathleen. (2007). 'CommentPress: New (Social) Structures for New (Networked) Texts'. *Journal of Electronic Publishing.* 3. `http://quod.lib.umich.edu/cgi/t/text/textidx?c=jep;view=text;rgn=main;idno=3336451.0010.305` (accessed 21 April 2010). .

Fortier, P. A. (1991). 'Theory, Methods and Applications: Some Examples in French Literature'. *Literary and Linguistic Computing.* 192–6.

Gabler, Hans Walter. (2010). '"Theorizing the Digital Scholarly Edition."'. *Literature Compass.* 2: 43-56.

Golder, Scott, Bernardo A. Huberman. (2006). 'Usage Patterns of Collaborative Tagging Systems'. *Journal of Information Science.* 2: 198-208. http://jis.sagepub.com/content/32/2/198.short (accessed 30 April 2010).

Hipp, Mason. (2008). '35+ Social Media Tools that Make Life Easier'. *Freelance Folder.* http://freelancefolder.com/35-social-media-tools-make-life-easier/ (accessed 30 April 2010).

Hoadley, Christopher M., Peter G. Kilner. (2005). 'Using Technology to Transform Communities of Practice into Knowledge-Building Communities'. *SIGGROUP Bulletin.* 1: 31-40 . [. http://doi:10.1145/1067699.1067705 (accessed 30 April 2010)] .

Burnard, Lou, Katherine O'Brien O'Keeffe, John Unsworth (eds.) (2006). *Electronic Textual Editing.* New York: MLA.

Dahlström, Mats. (2004). 'How Reproductive is a Scholarly Edition?'. *Literary and Linguistic Computing.* 1: 17-33.

Inman, James A., Cheryle Reed, Peter Sands (2004). *Electronic Collaboration in the Humanities*. Mahwah, NJ: Lawrence Erlbaum.

Irvine, Dean. '"Editing Archives / Archiving Editions"'. *Journal of Canadian Studies.* 2: 183-211.

Lancashire, D. Ian (1989). 'Working with Texts'. *IBM Academic Computing Conference.* Anaheim.

Manovich, Lev. (2001). *The Language of New Media*. Cambridge: MIT Press.

McGann, Jerome. (2004). '"Marking Texts of Many Dimensions."'. *Companion to Digital Humanities.* Susan Schreibman, Ray Siemens, John Unsworth (eds.). Oxford: Blackwell, pp. 198-217..

McGann, Jerome. (2002). 'Visible and Invisible Books: Hermetic Images in n-Dimensional Space'. *Literary and Linguistic Computing.* 2: 61-75.

McGann, Jerome. (2001). *Radiant Textuality: Literature after the World Wide Web*. New York: Palgrave Macmillan.

Ore, Epson S. (2004). 'Monkey Business, or What is an Edition?'. *Literary and Linguistic Computing.* 1: 35-44.

O'Reilly, Tim. (2005). 'What is Web 2.0?'. http://oreilly.com/web2/archive/what-is-web-20.html (accessed 4 May 2010).

O'Reilly, Tim, John Battelle (2009). 'Web Squared: Web 2.0 Five Years On'. *Web Summit.* San Francisco, CA. http://www.web2summit.com/web2009/public/schedule/detail/10194 (accessed May 10 2010).

Robinson, Peter, Hans Walter Gabler (eds.) (2000). 'Introduction. Making Texts for the Next Century'. *Literary and Linguistic Computing.* 15.1..

Robinson, Peter. (2000). 'The One Text and the Many Texts'. *Literary and Linguistic Computing.* 5.

Schreibman, Susan, Ray Siemens, John Unsworth (2004). 'The Digital Humanities and Humanities Computing: An Introduction'. *A Companion to Digital Humanities* Schreibman, Susan, Ray Siemens, John Unsworth (eds.). . Oxford: Blackwell.

Shillingsburg, Peter L. (2006). *From Gutenberg to Google: Electronic Representations of Literary Texts.* Cambridge: Cambridge University Press.

Shillingsburg, Peter L. (1998). *Resisting Texts: Authority and Submission in Constructions of Meaning.* Ann Arbor, MI: University of Michigan Press.

Shillingsburg, Peter L. (1986). *Scholarly Editing in the Computer Age*. Athens: University of Georgia Press.

Siemens, Ray, Cara Leitch (2009). 'It May Change my Understanding of the Field'. *Digital Humanities Quarterly.* 4. http://www.digitalhumanities.org/dhq/vol/3/4/000075/000075.html (accessed 15 April 2010).

Siemens, Ray, Christian Vandendorpe (2006). 'Introduction: Canadian Humanities Computing and Emerging Mind Technologies'. *Mind Technologies: Humanities Computing and the Canadian Academic Community.* Ray Siemens, David Moorman (eds.). Calgary: Calgary University Press, pp. xi-xviii.

Siemens, Ray (2005). 'Text Analysis and the 'Dynamic' Edition? A Working Paper, Briefly Articulating Some Concerns with an Algorithmic Approach to the Electronic Scholarly Edition'. *CHWP.* A.37/65. http://journals.sfu.ca/chwp/index.php/chwp/article/viewArticle/A.37/65 (accessed 12 May 2010).

Siemens, Ray (2002). 'Shakespearean Apparatus? Explicit Textual Structures and the Implicit Navigation of Accumulated Knowledge'. *Text: An Interdisciplinary Annual of Textual Studies 14* . Ann Arbor: University of Michigan Press.

Siemens, Ray (2001). 'Unediting and Non-Editions. The Theory (and Politics) of Editing'. *Anglia.* 3: 423-455.

Siemens, Ray, et. al. (forthcoming). 'Underpinnings of the Social Edition? A Narrative, 2004-9, for the

Renaissance English Knowledgebase (REKn) and Professional Reading Environment (PReE) Projects'. *Online Humanities Scholarship: The Shape of Things to Come.* Jerome McGann (ed.). Rice UP.

Tanselle, G.T. (1995). 'The Varieties of Scholarly Editing'. *Scholarly Editing* D.C. Greetham (ed.). . New York: MLA, pp. 9-32.

Terras, Melissa. (2009). 'Crowdsourcing Manuscript Material'. http://melissaterras.blogspot.com/2010/03/crowdsourcing-manuscript-material.html (accessed 2 March 2010).

Unsworth, John. (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 30 june 2009).

Wenger, Etienne. (2006). 'Communities of Practice'. http://www.ewenger.com/theory/ (accessed 29 Apr. 2010).

The Born Digital Graduate: Multiple Representations of and within Digital Humanities PhD Theses

Webb, Sharon
sharon.webb@nuim.ie
Department of History, National University of Ireland, Maynooth

Teehan, Aja
aja.teehan@nuim.ie
An Foras Feasa Research Institute, National University of Ireland, Maynooth

Keating, John
john.keating@nuim.ie
An Foras Feasa Research Institute, National University of Ireland, Maynooth

1. Introduction

This paper describes the methodology used in the creation of digital chapters and subsequent recreation of digital entities or objects derived, modified, transformed and visualised from XML encoded scholarship. It considers the changing function of traditional printed theses and how the use of technologies affects the representation and functions of graduate digital scholarship.

This paper is based upon the working methodologies of two PhD theses. Specifically, Webb's thesis examines the creation of factlets and subsequent visualisation of factoids, which inform not only the source information and encoding but also the development and completion of historical research outputs. These outputs, supported by XML, XQuery and factlets, demonstrate the use of digital technology as an essential feature of humanities research and its methodologies. Teehan's thesis reflects upon current digital representation models for pre-existing sources relevant to humanities research. Focusing on transactional, or functional, documents, it proposes a methodology for contextually modeling and XML-encoding those resources, using established software engineering and computer science paradigms such as Use Case analysis and UML modeling, which foreground the User. Both theses examine procedures and strategies for conducting humanities research using digital tools and applications. Thus, this paper is

central to a reflective and reflexive process resulting from, and in, the critical self-evaluation of the theses and their outputs.

Traditionally, research outputs codified as chapters or sections can be seen as the final manifestation of a PhD thesis and reflect the use of print or static technology. The functionality of these outputs varies according to different headings and ranges from literature reviews, general narrative and concept generation, to the development of structured arguments based on theory and source material, to the provision of essential referencing and bibliographic material. These functions are referred to as “generic characteristics of academic discourse” (Mingwei, 2010) in linguistic structural analysis. Chapter functionality represents and reflects the original research statement and provides the means to convey and articulate traditional scholarship within the medium of print. The use of XML, and XSLT, along with the provision of software libraries, creates a framework to add dynamic functionality to an otherwise static text. “Generic characteristics” (Mingwei, 2010) are encoded, which enable the use of the described framework.

This approach reflects the innate capability of the digital medium to layer extra functionality over the restricted functionality of printed works. Rather than creating just a single representation of scholarly output, the use of XSLT and software libraries generates and encourages a reflexive process between text, argument, narrative and source material.

These methods change reader and user activity - one user may be a reader while another may have access to an interactive environment. Different user roles and environments transform the user from a passive participant to an active one. The realisation of various use cases enables the user to do more than just read the text and this activity realises the importance of data reusability.

Figure 1 outlines the process involved in creating multiple representations of digital scholarship and will be used to detail the various stages involved in creating new digital objects based on specified use cases.

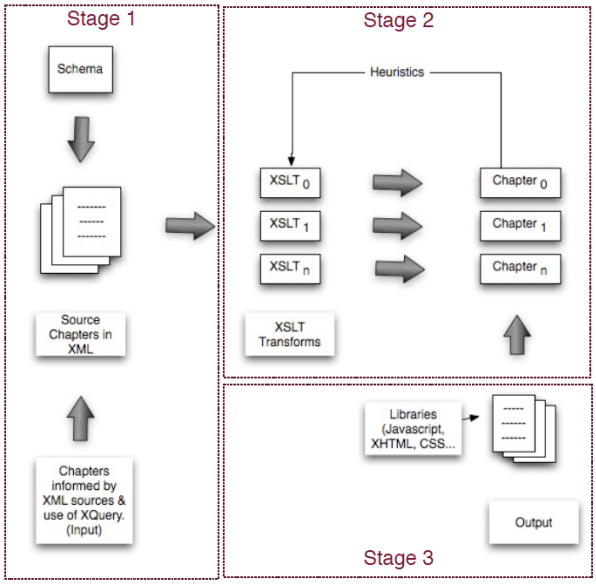


Figure 1: The stages of creating new research outputs, various chapters are defined by various use cases.

2. Stage 1 - Defining the Use Cases, Creating the Model, Encoding the Source

Text is innately encoded with semantics and functionality and each chapter or section in a piece of scholarship establishes or conveys various essential processes in the life of a text. These processes consist of deduction, concept development, narrative, consequence, etc., and it is these “lexical relations” (Eggins, 2004) within a text which develop specific research statements. Other bibliographic properties of a section are concrete rather than abstract and provide essential functionality e.g. references, footnotes, paragraphs and titles.

Despite these multiple perspectives, transformation of a born-digital text (a thesis) into both the print and digital media relies upon the existence of a single, defining text-model. Figure 1 shows the process involved in creating new research objects. The first stage makes possible all subsequent processes; creating a unifying model allows the generation of XML schema and subsequent XML encoding in order to manifest the new research objects (the various chapters in a research thesis). The model is driven by specific Use Cases such as the production of both a static printed version of the text, and an interactive digital version.

The source chapters are encoded at the final stages of the research process, rather than during the writing process. The model considers both presentation properties (chapter, paragraph, section), which allows for transformations specifically for

presentation purposes, and semantic properties which encode the “textual semantics” (Eggins, 2004) of the text, its logical class (Teehan, 2010). This approach makes the text reusable and ensures “a single lexical can function very differently” (Landow, 2006) in different environments.

The model is translated into a schema which allows us to mark up the content of the scholarship, including narrative which in historical research pertains to ‘logical’ rather than ‘ideological’ content. We view narrative as the logical information contained in the text that contributes to a narrative of the past (Coffin, 2002). The encoding of dynamic narrative and data supports the creation of new research outputs as non-linear components derived from the text.

3. Stages 2 and 3 - Realising Use Cases

Stage 2 and 3 are the realisation of the various Use Cases. The XSLT transforms and software libraries are templates from which different text from different sources can be modified and transformed, in effect creating a suite of tools.

These various macros are supporting tools for manifestations of a text. Our encoded texts depict the various functions embedded in standard print theses, but augment those capabilities for these born-digital theses. Here, two specific Use Cases address (i) the creation of a dynamic bibliographic referencing model, and (ii) the context-dependent presentation of boundary objects.

A referencing model in XSLT can automatically create a dynamic bibliography for a chapter with features including “intertextual links” (Samraj, 2008) between the text and source material. Software libraries can be used to support the innate variability of a boundary object, which is defined as an object with user dependent functionality and meaning (Thomas). Thus, depending on the User’s activity and perspective, the presentation of the boundary object will change; for instance, a table diagram, static in the print version, may become interactive within a digital context.

These low-level Use Cases support our higher level one; dynamic creation of static or interactive versions of a base text-model. The print model transforms the original text to a print ready text, and can account for various institutional templates. Embedding references to the various primary sources used in the XML encoding instructs an XSLT to create a hypertext of linked resources and creates “intertextual links” (Samraj, 2008) and boundary objects for user interaction between the narrative and various digital objects within the digital medium. Figure 2

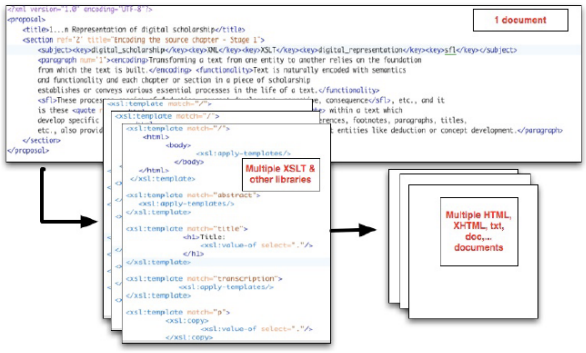


Figure 2: Text encoding of this proposal and XSLT transforms

4. Conclusion

This paper demonstrates the process and production of support tools for digital scholarship, and how the creation of appropriate templates can make manifest various representations of Digital Humanities PhD theses from a single model. The Use Cases are reliant on the ability of the encoding and the schema to encapsulate both the functions of the text and the various transformations and software libraries. Figure 2 demonstrates the interactions between the encoded text, the transformations and the outputs.

Current research students in Digital Humanities constitute a newly ‘born digital’ generation, the nature of whose outputs differs markedly from earlier generations. Reflections on this changing process should also include an analysis of new methods and techniques to create dynamic scholarship. The encoding of the final phase in a PhD thesis allows scholarship to be reused, modified, visualised and transformed, allowing for greater distribution and accessibility of digital scholarship. Thus the dissertation, in its multiple representations, can not only remain central to the discipline of Digital Humanities but shape its future development.

References

Mingwei, Zhao, Yajun Jiang (2010). 'Dissertation acknowledgements: Generic structure and linguistic features'. *Chinese journal of applied linguistics* 33 (1) 94-109.

Eggins, Suzanne (2004). 'An introduction to systemic functional linguistics'. *Continuum.* .

Teehan, Aja, John G. Keating (2010). "A digital edition of a Spanish 18th- century account book, Part 1: User-driven digitisation". *Jahrbuch für Computer-Philologie* 10 169-18. http://computerphilologie.tu- darmstadt.de/jg08/keating1.html.

Landow, George P (2006). *Hypertext 3.0, Critical theory and new media in an era of globalization.* The John Hopkins University Press.

Coffin, Caroline (2002). 'Constructing and giving value to the past: an investigation into secondary school history'. *Genre and institutions, social processes in the workplace and school.* Frances Christie (ed.). London: J.R. Martin.

Samraj, Betty (2008). 'A discourse analysis of master's theses across disciplines with a focus on introductions'. *Journal of English for academic purposes.* 55-67.

Thomas, Robyn, Sargent, Leisa D, Hardy, Cynthia. *Power and participation in the production of boundary objects.* http://www.s-as-p.org/files_news/Thomas%20et%20al%20-%20power%20and%20participation%20in%20the%20production%20of%20boundary%20objects.doc.

Computational Analysis of Gender and the Body in European Fairy Tales

Weingart, Scott
scbweing@indiana.edu
Indiana University, United States of America

Jorgensen, Jeana
jeanaj@gmail.com
Indiana University, United States of America

This paper presents preliminary results on using computational analysis to understand the representations and constructions of gender and the body in fairy tales. While scholarship on contemporary fairy tales utilizes various cutting-edge theories, ranging from postmodern narrative to feminist theories of gender performance (Bacchilega 1997, Benson 2008, Smith 2007, and Tiffin 2009), little of the research on canonical fairy tales or oral folktales incorporates these recent theories. Additionally, folkloristic research on fairy tales, whether contemporary or traditional, would benefit from incorporating computational methods such as network analysis. These methods allow scholars to test their theories more quickly and empirically.

Our research utilizes nearly three hundred canonical fairy tales and oral folktales, deemed canonical because they are from well-known collectors such as the Brothers Grimm, or because the tales are examples of well-known plots spanning time and space in Europe (such as “Snow White” and “Cinderella”). We combine textual and network analysis with discipline-specific expert oversight for a large-scale, theoretically informed discussion on gender and the body that would not be possible without both in tandem. A feminist critique of fairy tales is predicated upon the notion that fairy tales construct and represent bodies differently according to gender, yet no studies have shown whether this difference actually exists in canonical tales, or have addressed what this difference would mean for studies of cultural values and narrative strategies (Bottigheimer 1987, Haase 2004, and Stone 2008). Computational analysis of how bodies and body parts are depicted in the text provides empirical evidence against which this and other aspects of feminist theory can be tested.

Humanities scholars have already established a vast theoretical and methodological framework for interpreting texts, and they ought to be able to view their data in the context of those theories developed within their disciplines. This study combines traditional critical analysis with computational tools in an attempt to utilize the best of both worlds.

2. Data

Our analysis uses a hand-coded database representing a geographically and temporally diverse sample of tales. Careful attention was paid to the tale tellers and collectors for further study of the context in which bodies are depicted.

Fairy tales as a genre span oral, communal performances and literary, single-author renditions. In order to represent this spectrum, our database tracks specific references to bodies in six tale collections. We collected 13 data points from nearly three hundred tales (Tale, Collection, Author, Teller, Collector, Year of Writing/Collecting, Year of Publication, Tale Type, Region, Original Language, Gender of Teller/Writer, Gender of Collector, Gender of Editor) and categorized another 14 data points for every mention of a body in each tale, some evident in the texts (Noun, Adjective, Surrounding Text, Page Number, Gender, Young/ Old, High/Low, Quoted Speech, Skin Tone) and some requiring interpretation (Positive/Negative value, Grotesque, Violence, Nudity, Move).

The database variables were chosen in light of pre-existing work on structure and theory, creating a layer of interpreted data that would not be found in full-text analysis alone. The “Tale Type” classification system gives tale plots numbers so that their transmission can be traced as tales migrate across linguistic and national boundaries. This is what allows us to generalize about the worldview contained within the tales, as the same plot with variations occurs between multiple ethnic groups. The concept of “Moves” breaks up tale plots into 5 distinctive plot pieces based on folkloristic theory of how tales are structured.

3. Methodology

We use co-occurrence and vector analysis to explore the database. Each field is compared against several others in order to find correlations. For example, “beautiful” may only be referenced with young women, or old rich men may only appear in tales from certain tellers. Using dimensionality reduction, we can find which body parts tend to be discussed in tandem in various situations. We also explore how the representations of bodies change throughout the plots of tales using Bengt Holbek’s “Moves.” Holbek built on the work of Vladimir Propp, who identified the most important plot points in sequence that could occur in a fairy tale (31 points, or “functions,” total) (Holbek 1998). Holbek condensed Propp’s functions into five “Moves,” or clusters of thematic actions that move the tale’s plot forward.

Finally, networks of database data are generated and analyzed to test the hypothesis that fairy tales construct bodies differently according to gender. This analysis serves both as empirical evidence to test a theory and also as an exploratory tool, revealing possible correlations or links between body representations that are not immediately apparent in the texts.

Folklorists approach fairy tale interpretation in many ways: ethnographic approaches seek connections between the taletellers’ lives and the tales’ content; historical approaches search for and analyze the origins and diffusion of tales; structural analyses seek to understand the underlying narrative of the tales; psychological approaches search for latent meaning in the tales; and feminist and sociohistorical approaches interpret the meanings of the tales as they relate to, convey values from, and inculcate values of the social world. Feminist scholars have been particularly active in critiquing the normative beauty scripts and gender roles promoted in fairy tales. This study investigates how gender roles are constructed and situated in fairy tales, which is why we encoded categories to investigate links between gender, age, and social position, as well as where in the tale’s structure these social values are relevant. We also hope to obtain information about how female and male bodies are valued differently, hence the relevance of variables like “grotesque”.

4. Preliminary Results

Second-wave feminists such as Simone Beauvoir developed the notion of the universal masculine perspective, the idea that in Western culture, the public, unmarked, assumed universal position is in fact specifically male. Our data supports this assertion in terms of female bodies being marked within fairy tales, but we also believe that the same principle applies to the age of bodies. Youthful bodies are assumed to be the unmarked universal category in fairy tales.

The descriptions of old bodies are strikingly polarized: old people are more likely than young to be described as evil or good, wicked or wise. These findings suggest that old bodies must be differentiated in fairy

tales, because they are no longer in the supposedly universal category of youth. Old bodies are qualified with more descriptions in order to give audiences a sense of who these characters are, since they don't fall into the category of the youthful protagonist, with whom listeners are supposed to easily identify. If the bodies in fairy tales had the same meanings across age and gender, we would have seen a proportional relationship between the number of references to types of bodies, and the number of adjectival descriptions attached to each. However, the data shows that young men are associated with adjectival descriptions less frequently than any other type of body. Old women, in contrast, are associated with adjectival descriptions more than any other grouping. Further, a wider variety of adjectives are used to qualify old bodies than young compared to the proportion they are mentioned. Figure 1 shows a sample of which adjectives are associated with mentions of various bodies.

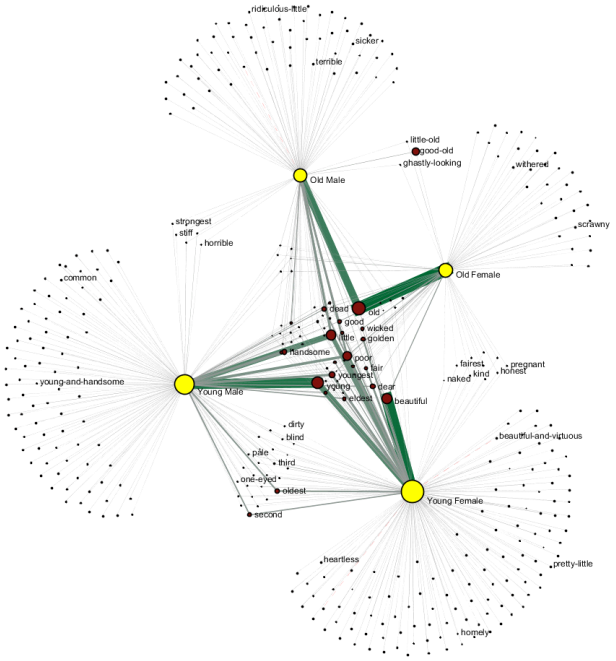


Figure 1. Lines are drawn between adjectives (red) and the body-types they modify (yellow). Node size represents word usage counts and edge thickness represents frequency of co-occurrence.

5. Goals

Our method of layering computational analysis atop previous theoretical research can be used as a template for further studies, especially those of other folk narrative genres like legends or ballads. Some of the most intriguing questions in folklore research pertain to how verbal expressive genres relate to the lived experiences of their performers—and a method for digitizing and interpreting these texts could yield valuable insights.

As digitization is interpretation (Tarte 2010), it is necessary to be especially careful and theoretically-grounded when choosing variables and selecting exactly what data will populate the fields. The scholar must also decide the most fruitful analyses to run on the data available. These studies ought to also include computational analyses that are not linked to previous critical theories (like word frequency or co-occurrence), however, in order to check against biases which might creep into variable choice. The ultimate goal is to turn well-researched, theoretically sound scholarly observations into machine actionable data which can be analyzed to test the scholar's hypotheses and open the door for future studies.

References

Bacchilega, Cristina (1997). *Postmodern Fairy Tales: Gender and Narrative Strategies*

Bakhtin, Mikhail (1984). *Rabelais and His World.* .

Benson, Stephen (ed.) (2008). *Contemporary Fiction and the Fairy Tale*

Blei, David, Ng, Andrew Y., Jordan, Michael I., Lafferty, John (January 2003). *Latent Dirichlet Allocation. Journal of Machine Learning Research.* .

Bottigheimer, Ruth (1987). *Grimms' Bad Girls and Bold Boys: The Moral and Social Vision of the Tales*

Haase, Donald (2004). *Feminist Fairy-Tale Scholarship. In Fairy Tales and Feminism: New Approaches.* Donald Haase (ed.). .

Holbek, Bengt (1998). *Interpretation of Fairy Tales*

Muhawi, Ibrahim, Sharif Kanaana (1989). *Speak, Bird, Speak Again: Palestinian Arab Folktales*

Propp, Vladimir (1968 [1928]). *Morphology of the Folktale.* .

Roberts, Warren (1994 [1958]). *The Tale of the Kind and Unkind Girls*

Russo, Mary (1994). *The Female Grotesque: Risk, Excess, and Modernity*

Smith, Kevin Paul (2007). *The Postmodern Fairytale: Folkloric Intertexts in Contemporary Fiction*

Stone, Kay (2008). *Some Day Your Witch Will Come*

Taggart, James M. (1990). *Enchanted Maidens: Gender Relations in Spanish Folktales of Courtship and Marriage*

Tarte, Ségolène M. (2010). *Digitizing the Act of Papyrological Interpretation: Negotiating Spurious Exactitude and Genuine Uncertainty. Digital Humanities 2010*

Tiffin, Jessica (2009). *Marvelous Geometry: Narrative and Metafiction in Modern Fairy Tale*

Zipes, Jack (1994). *Spreading Myths about Iron John. Fairy Tale as Myth, Myth as Fairy Tale.* .

The UCLA Encyclopedia of Egyptology: Lessons Learned

Wendrich, Willeke
wendrich@humnet.ucla.edu
UCLA

The UCLA Encyclopedia of Egyptology (UEE) is a digitally born online publication which provides users with several interfaces to access and reproduce content. Supported by several grants from the National Endowment for the Humanities, the UEE is a highly mediated, peer reviewed information resource on Egyptian history, art, archaeology, geography and language, in which authors selected by an editorial committee are commissioned to write substantial articles with thorough bibliographies and web links. Articles will be regularly updated and previous versions of the text and other assets will remain available throughout the lifetime of the resource, which in principle is built for digital eternity. This time scale may prove not to be of the same longevity as the preservation of ancient Egyptian cultural heritage, but is nevertheless fitting the mindset of scholars who routinely deal with objects of 4000 years old.

The UEE is an English language resource, while all head words or entry titles are translated also in Arabic, French and German. The English abstracts are also translated in Arabic and a standard feature of each article. The content of the UEE is available in two forms: the Open Version makes use of eScholarschip, the online publication platform of the University of California. Articles are presented in alphabetic order of the titles, and can be downloaded as PDF (http://escholarship.org/uc/nelc_uee). The UEE Full Version provides a much more sophisticated platform, where users can access information through a wide range of searches, either based on the underlying subject structure, article links, metadata, or through an interactive time-map, which provides access to articles, images and 3D VR reconstructions which refer to the same area, the same time period or both. The granularity of the time map encompasses regions (using either modern or ancient subdivisions), ancient sites, or particular features of the latter, such as a specific gate way, or altar. At present the Full Version is not yet publicly online, but will be moved from development to production in the near future. Information on the URL will be provided on the UEE project development website at http://uee.ucla.edu.

The presentation will focus on the many lessons learned while developing the project, including the workflow of all the tasks which are literally performed "behind the screen". Since this is an international project with editors in the United States, Canada, France, Belgium, Great Britain, and Egypt, with authors as well as peer reviewers from all over the world, the project relies on a large number of disparate web services, which are partly for free, partly subscription based. The commissioning phase is tracked through a number of online spread sheets through Google Docs. Authors receive an invitation by email, which provides them with the scope of the entry, a document with clear indications of what should (not) be included, in order to avoid duplication with related articles. Once an article has been submitted, the peer review platform provided by eScholarship, is used, which enables tracking and automated prodding for authors and peer reviewers. Since many of the authors are non-native speakers of English, the UEE offers a substantive copy-editing service, which streamlines the terminology, spelling and links of the articles. The project coordination, which involves communication on the progress of the extensive mark-up in TEI which is the next phase of bringing an article to online publication, makes use of the commercial project management software BaseCamp (http://37signals.com/).

An important point of discussion is the digital and financial sustainability and the different solutions the UEE has proposed to enable the project to expand and be constantly renewed, which is perhaps the greatest asset of an online resource. The history of the venerable printed predecessor of the UEE, the Lexikon der Ägyptologie (published from , shows that bibliographies are outdated in five years after publication, while Egyptological scholarship begins to be outdated in approximately 20 years. Authors are therefore asked to provide twice an update of their article, and after that potentially a new entry will be created, because the development of the discipline is not only reflected in article content, but also in the structure of the resource, the selection of entry titles, and the emphasis on particular sub fields. The strict version control has, therefore, an added benefit: over the course of time the UEE will become effectively a history of Egyptological thought and methodology.

This requires, however, that the digital content remains available, and that the editorial process will keep on running, two very demanding conditions. As for digital stability: all assets of the UEE are housed in the UCLA Digital Library, and are accessed from a front-end server which is at present housed at UCLA's Academic Technology Services. This agreement does not necessarily guarantee digital preservation, because the libraries, at the forefront of digital repository preservation as they are, are also faced with the enormous costs and practical problems of digital preservation. The awareness of the problem is, however, a considerable part of the solution. The financial sustainability is an enormous conundrum. Users have come to expect free, high quality content, academics celebrate the virtue of open access, and yet to build a high quality resource comes with very real costs that need to be covered. The presentation will outline some of the avenues explored to ensure that the UEE will have a long prosperous life.

Possible Worlds: Authorial Markup and Digital Scholarship

Wernimont, Jacqueline
jwernimo@ScrippsCollege.edu
Scripps College

Flanders, Julia
Julia_Flanders@brown.edu
Women Writers Project, Brown University, United States of America

Alan Liu's "Imagining the New Media Encounter" (2008) calls for "a poiesis of digital literary studies" through which we can renegotiate the relationships between new and old media as productive encounters rather than as something other than "conversion" encounters. Liu helps us open up a critical space in which to rethink how the problematic notion of "conversion," with its implications of oppositional media, complete transformation, and religious fervor, shape our understandings of related pairs: writing and encoding, mimesis and creation, imagination and simulation. We suggest that our understanding of text markup is closely implicated in our reimagination of writing, and that the modes of modeling suggested by possible worlds scholars may destabilize our understanding of mimesis and its role in both literary composition and text markup. In short, we propose considering markup as a "world-constructing" (Doložel 1997) form of discourse.

Our starting point of reference is the long-standing conceptual tension within the markup (and especially the TEI) community between two models of markup. The first is rooted in mimesis and surrogacy: the domain of transcriptional and editorial markup. The second is more concerned with meaning creation and the domain of annotation, interpretation, authoring. These two models have different textual commitments and establish different relationships between text, markup, reader, and encoder. In the first, the encoder uses markup to transact a connection between a text and a reader; it is understood that the markup is non-transparent, but its role is to communicate about the text and about its own role as transmission medium, so that the reader can (to the greatest extent possible) apprehend some truth of the text. The primary commitment, the goal of the exercise, is for the reader to have access to the text (that is, to some textual artifact that pre-existed the markup relationship). Some form of this approach is extremely common in current applications of the TEI Guidelines: for thematic research collections, scholarly editions, linguistic corpora, oral histories, digital archives, and the like.

The second model, though much less common in practice, is of great importance theoretically as a counterpoise to the first, and its importance has been shadowed forth by a number of key interventions during recent years. Theorists like Renear (2000) and McGann (2004) in very different ways have suggested that the performative and illocutionary qualities of markup bear close scrutiny. Sperberg-McQueen and Huitfeldt (2000) explored how markup represents meaning (and by extension, suggest a shift of emphasis onto markup itself as a meaning-bearing system, apart from the text it marks). Flanders (2006) and Flanders and Fiormonte (2007) have turned attention to the rhetoric of authorial markup and to its significance for scholarly communication, thinking of markup as a discourse that is situated at the boundary of production and reproduction. In this more "authorial" model, the encoder uses markup to transact a connection between *herself* and a reader *that concerns* a text. The role of markup is to instantiate, to bring into communicative reality, the encoder's ideas and beliefs about a textual ecology that is oriented towards a particular textual artifact but is not limited to representing that artifact. Rather, the markup may represent a much broader context of interpretation, related information, and argumentation for which the text itself is only the catalyst or point of inspiration. The most common examples in the present day include annotation, "interpretive" markup such as the association of themes and keywords with spans of text (e.g. using the TEI @ana and *interp* mechanism), and the creation of new documents such as articles using an XML markup language as an authoring system. But these examples do not really give us a field within which to consider what —in a radical sense— we might mean by "authorial markup", or to pursue the full critical pressure of McGann's challenge to the markup world: "No autopoietic process or form can be simulated under the horizon of a structural model like SGML" (McGann 2004, 201-2).

The way these questions are framed within the digital humanities—as an opposition between mimetic representation and presentation on the one hand, and generative or creative authorship and interpretation on the other—has a correlate in literary history. As Doležel argues (1997), possible world texts, with their "world-constructing" features, contrast with the strongly mimetic and material world truth claims of "world-imaging" texts. We suggest that an exploration of a

"possible worlds" ontology of non-mimetic discursive modes may offer a critical vocabulary for thinking about the relationship of authorial markup to other encoding models. It can also help us describe how authorial markup might leverage the formal tools of a structural model in order to enact a generative or poetic mode of markup.

We focus in particular on the authoring, interpretation, and markup of texts from the early modern period that concern themselves with precisely this domain. While Gottfried Leibniz nominally inaugurated discussion of possible worlds (in *Essays on the Goodness of God, the Freedom of Man and the Origin of Evil*, 1710), early modern writers from Thomas More and Sir Philip Sidney to René Descartes and Margaret Cavendish were concerned with the ability of the written word or number to, as Descartes put it, write about that "which does not actually exist…but is capable of so doing" (Descartes, Writing, 332). Such propositional discourse refers less to a verifiable "real" than to a set of possibilities without a fixed ontological status. Instead, as Ruth Ronen suggests, "their state of being is confined to what meaning-units of the text reveal" (Ronen 1994, 98-9). The early modern romance in particular (of which Cavendish is a major exemplar) is the single prose genre most invested in explicit exploration of the mimetic/poetic distinction and in which narrative structures themselves enact the generative logic of world-construction. Cavendish herself argues in her 1667 *A New World Called the Blazing World*, that her "romancicall" tale offers readers a model by which "they may create worlds of their own, and govern themselves as they please" (Cavendish 225). Her model of romantic poiesis draws on a poetic tradition exemplified by Sidney's *Defense of Poesie* (1579), which refutes directly the critique that poiesis is a form of feigning, by suggesting an epistemological and ethical role for generative authoring.

When we turn from Cavendish's own authoring process to our own use of markup as a means of representing and interpreting her text—and of creating a new work of scholarship of which we are the authors—these arguments have a double impact. Within the world of an encoded text, the "meaning-units" include (for our purposes, most significantly) the markup itself. The markup itself becomes a world-generating mode of knowing that must carry several registers of meaning arising from different kinds of scholarly agency. There are also important questions which we will address in the full paper about how we can anchor and access the domain of meaning such a world establishes.

Following Saul Kripke's assertion that "possible worlds are stipulated, not discovered by telescopes," we see in both Cavendish's authoring and our reading and encoding of her text a series of generative or poetic "stipulations" that carry both epistemological and ethical implications (cited in Doložel 1998, 787). We stipulate that a given structure can be characterized as a poem, or that a new paragraph begins in this place. Because there is such agreement (e.g. on disciplinary grounds) about how we read and name literary structures, this stipulation reads like a statement of fact, but if we consider more unfamiliar genres (or genre-resistant texts, like the confounding recursive narratives of Mary Wroth's *Urania* or entries in commonplace books), the contentiousness of the assertion becomes more apparent. Possible worlds theory elaborates on what such stipulations accomplish, and on the terms in which we can understand the truth-value of what is created thereby. In the full version of this paper, we will consider how schemas may operate as another way of modeling the possible worlds of texts and genres, and also how possible worlds theory may help us envision more radically authorial forms of markup that may push our ideas of text encoding as scholarly communication into new terrain.

References

Cavendish, Margaret, Duchess of Newcastle (1666). *Observations on Experimental Philosophy. To Which is Added, The Description of a New World, Called the Blazing-World.* .

Descartes, René (1637) (2001). *Method, Optics, Geometry, and Meteorology.* Paul J. Olscamp (ed.). .

John Cottingham, Robert Stoothoff, Dugald Murdoch, Anthony Kenny (eds.) (1991). *The philosophical writings of Descartes, Volume 3.* .

Doležel, Lubomir (1997). *Heterocosmica: Fiction and Possible Worlds.* .

Doležel, Lubomir (1998). 'Possible Worlds of Fiction and History'. *New Literary History.* 29.4: 785-809.

Flanders, Julia (July 2006). 'The Rhetoric of Performative Markup'. *DH2006.* .

Flanders, Julia, Domenico Fiormonte (June 2007). 'Markup and the Digital Paratext'. *DH2007.* .

McGann, Jerome (2004). 'Marking Texts of Many Dimensions'. *A Companion to Digital Humanities.* Susan Schreibman, Ray Siemens, John Unsworth (eds.). 198-217.

Poovey, Mary (2008). *Genres of the Credit Economy.* .

Renear, Allen (2001). 'The Descriptive/Procedural Distinction is Flawed'. *Markup Languages: Theory and Practice.* 2.4: 411-420.

Ronan, Ruth (1994). *Possible Worlds in Literary Theory.* .

Interedition: Principles, Practice and Products of an Open Collaborative Development Model for Digital Scholarly Editions

van Zundert, Joris
joris.van.zundert@huygensinstituut.knaw.nl
Huygens ING - Royal Netherlands Academy of Arts and Sciences

Middell, Gregor
gregor.middell@uni-wuerzburg.de
Universität Würzburg, Lehrstuhl für Computerphilologie

Van Hulle, Dirk
dirk.vanhulle@ua.ac.be
University of Antwerp, Centre for Manuscript Genetics

Andrews, Tara L.
tara.andrews@arts.kuleuven.be
Katholieke Universiteit Leuven, dept. of Greek Studies

Haentjens Dekker, Ronald
ronald.dekker@huygensinstituut.knaw.nl
Huygens ING - Royal Netherlands Academy of Arts and Sciences

Neyt, Vincent
vincent.neyt@ua.ac.be
University of Antwerp, Centre for Manuscript Genetics

---

1. Short Paper Abstract

In October 2006 a small group of developers of tools for digital textual scholarship, gathered under the leadership of the Huygens ING, concluded that there was an urgent need for a more collaborative approach to digital tool development in the humanities. Several problems plagued the field: high duplication of effort, a shortage of quality software development, poor exchange of development and methodological knowledge, institutionalized development, and consequent problems of non-sustainability and obsolescence. In 2008 this initiative became a formal European funded project, COST Action IS0704 'Interedition'. Since then, Interedition

has been improving cooperation and fostering interoperability in tool development for digital textual scholarship. [1] This paper will reflect on the first results that have emerged meanwhile.

We have identified four significant obstacles to the widespread use of digital tools by humanities scholars. First, applicability of tools to humanities research is often lacking: in many cases the tools are not tools researchers need. Second, development capacity within the humanities is extremely limited compared to scientific fields. Third, the sustainability of these tools, in terms of their ongoing support and maintenance requirements, is historically very low. Finally and perhaps most critically, tool and infrastructure availability and development is highly institutionalized: scholars outside large research projects or institutions often find themselves unable to take advantage of the development work that might otherwise benefit them substantially.

The researchers and developers in the Interedition project have come up with an approach to the problems of applicability, availability, and sustainability that revolves around the concept of 'microservices'. A microservice is, ideally, a very small application whose functionality is available over the Web by means of a lightweight protocol (e.g. REST). Microservices can be used programmatically in conjunction with other such services, in multiplicative combinations, to answer the individual need of any research project. Microservices share a set of deceptively simple principles: they are cheap and fast to develop; cheaper and easy to maintain; address very specific needs (that are shared between many researchers); implement simple protocols that are easier to reuse and exchange; can be combined to create larger workflows useful to the individual scholar. This idea is not new: already service-based architectures such as SEASR/MEANDRE, and commercial services such as Yahoo! Pipes, feature a 'modular' model, and such architectures are increasingly considered to be the best approach for software development within humanities research.[2] However, given the variety of contexts and environments in humanities research, it is critical that tools not be tied to a single unifying infrastructure, which is a drawback of systems such as SEASR or Yahoo! Pipes. If we are to make maximally efficient use of humanities development capacity, any scholar or developer should be able to contribute his or her work to other projects with a minimum of effort; this means that we cannot insist upon a standardized infrastructure, platform, computing language, or even data format. Moreover, there is no guarantee that any such single infrastructure will be indefinitely available; the infrastructure thus becomes a single point of failure.

Over the past two years, Interedition has been putting its theory to the test by exploring effective development methods for tools for text criticism and literary analysis, using the very limited resources available through the COST framework. The result is a open-source development 'collective' cooperating in 'bootcamps'. Cooperation does not require everyone to have the same research goal, nor need they agree on the 'right' way to approach textual scholarship. The resulting tools are furthermore not tied to or centralized in any single institution. The combined efforts of the participants become a collection of microservices, varying widely in implementation language, platform, hosting provision, etc., according to the resources and expertise available to the individual participant. The only technical requirements placed on any microservice are that it be web-accessible to other services using a REST-like protocol, and that information on its input requirements and output results be available via an HTTP GET request.[3] Thus, even with a wide variety of implementation practice, small microservices can be built up to perform tasks in a larger workflow for 'real' and varied research purposes.

Current 'live' microservices are open source, which is crucial both to the proposed model of cooperation for research and development, to intellectual and scientific transparency, and to sustainability. Many of them are hosted on free cloud computing infrastructures (e.g. Google App Engine or Heroku[4]) for increased availability and reliability.

As a proof of principle Interedition chose to implement as a microservice architecture a tool commonly wanted for textual scholarship—collation of text witnesses—and worked to design a set of web-based services that could improve upon existing technologies. Designing collation software for textual editors that would substantially improve upon existing tools like the NINES project's JUXTA[5] and Peter Robinson's COLLATE[6] has been a significant challenge, and one that has allowed the team to thoroughly test their development model and architectural approach with an important problem which, if solved, would immediately benefit a broad community of scholars. The result is CollateX[7], developed as a technical and methodological successor to COLLATE, which reached the end of its supported life around 2007. We will describe CollateX's decentralized development within the Open Source community and how 'bootcamps' brought together an international team of developers and domain

experts for requirements analysis, coordination of development efforts and collaborative work on the code base. We will also explain how Interedition's design principles were implemented by splitting the collation process into functional tasks, each of which can be implemented with a variety of small and well-defined software services; these services are then loosely coupled to produce "the simplest solution that could possibly work" for the different use cases to be addressed. Three clearly separate functional tasks—input tokenization, the alignment itself, and analysis or visualization of the results—were identified and implemented as microservices.

This 'microservice model', as exemplified by CollateX, has another important advantage: it fits well with current directions of thought on the form and function of future digital scholarly editions. Rather than simple and static online republication of books, it is likely that future editions will be dynamic open-ended research environments[8] composed of smaller scholarly components from many different sources. It is an ultimate goal of Interedition to leverage this microservice architecture to facilitate such 'distributed editions'.

2. Poster Abstract

This poster presentation will demonstrate in detail the technical aspects of Interedition's 'microservices model' for interoperability. The poster will serve as the technical annex to the short paper on Interedition and its development principles, and will describe in detail the working of the various components that make up CollateX, Interedition's foremost proof-of-concept implementation. At the poster presentation CollateX will be running on a laptop to demonstrate its capabilities.

We will also showcase some digital humanities projects that are benefiting already from Interedition's approach to tool building. Examples include TILE and T-PENN, both of which focus on the common task of transcribing images of text and aligning this transcription with the associated regions in the image. When their commonalities are examined, a similar set of core services might be deduced and a robust set of modular services designed to improve upon the advances these tools have already made. We will take a more in depth look at the Beckett Digital Manuscript Project on which the Centre for Manuscript Genetics (University of Antwerp) is collaborating with the Huygens ING. This project raises the question within the framework of Interedition of how the architecture of a digital archive containing modern manuscripts can be designed in such a way that users can autonomously collate textual units of their choice with the help of Interedition's collation web service and thus decide for themselves how this digital architecture functions – as an archive, as a genetic dossier, or as an edition.

As a consequence of developments in digital scholarly editing, the strict boundary between digital archives and electronic editions is becoming increasingly permeable, resulting in a continuum rather than a dichotomy. Usually, archives are distinguished from editions because the latter offer a critical apparatus. Of all the interoperable tools developed within the Interedition framework, the collation module has the special merit that it can enable any user to transform a digital archive into an electronic edition.

From the vantage point of editorial theory, this development has interesting consequences regarding the scholarly editor's role, whose focus may shift from the collation to a more interpretive function. In this way, the integration of a collation tool may be consequential in terms of bridging the gap between genetic criticism and textual scholarship. From the perspective of editorial practice, the application of CollateX is still at an experimental stage, but it already shows that the modular approach used by Interedition has the potential to be useful both to the specialized field of digital scholarly editing and to a more general audience.

References

Atkins, D. et al (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* NSF.

Cohen, D. et al. (March 2009). *Tools for Data-Driven Scholarship: Past, Present, Future. A Report on the Workshop 22-24 October, 2008, Turf Valley Resort, Ellicott City, Maryland.* . `http://mith.umd.edu/tools/final-report.html`.

European Union (2010). *Riding the wave. How Europe can gain from the rising tide of scientific data. Final Report of the High Level Expert Group on Scientific data. A submission to the European Commision.* Italy.

Robinson, P. (2010). 'Electronic Editions for Everyone'. *Text and Genre in Reconstruction. Effects of Digitalization on Ideas, Behaviours, Products and Institutions.* Willard McCarty (ed.). Cambridge, UK: OpenBook Publishers.

Notes

1. Van Zundert, J. et al.: Memorandum of Understanding (MoU) for the implementation of a European Concerted Research Action designated as COST Action IS0704: An interoperable supranational infrastructure for digital editions (Interedition). Brussels, Belgium, 2007. http://snipurl.com/1dl4v1, accessed 30 October 2010.
2. Küster, M. W., Ludwig, C., and Aschenbrenner, A.: 'TextGrid as a digital ecosystem', IEEE DEST 2007, 21.-23. Cairns, Australia, 2007. SEARS/MEANDRE: http://seasr.org/meandre/documentation/architecture, accessed 30 October 2010; http://seasr.org, accessed 30 October 2010. Yahoo! Pipes: http://pipes.yahoo.com, accessed 30 October 2010.
3. http://gregor.middell.net/collatex/api/collate, accessed 30 October 2010
4. http://code.google.com/appengine/, accessed 30 October 2010; http://heroku.com/, accessed 30 October 2010.
5. Juxta. Collation software for scholars: http://www.juxtasoftware.org/, accessed 31 October 2010.
6. Robinson, P.: 'Collate: A Program for Interactive Collation of Large Textual Traditions', in N. Ide and S. Hockey (eds.), Research in Humanities Computing. Oxford, 1994, pp. 32–45.
7. http://collatex.sourceforge.net, accessed 30 October 2010
8. Boot, P. and Van Zundert, J.: "The Digital Edition 2.0 and The Digital Library: Services, not resources." In: Knoche, M., Mittler, E. et al. (eds): Bibliothek und Wissenschaft. Wiesbaden, Germany. (Forthcoming); Robinson, P.: http://computerphilologie.uni-muenchen.de/jg03/robinson.html, accessed 30 October 2010.

Posters

Digital Collections at Duke University Libraries

Aery, Sean
sean.aery@duke.edu
Duke University

Sexton, Will
will.sexton@duke.edu
Duke University

The digitization of primary sources for humanities research marks one of the important ways that the emergence of a digital culture has transformed libraries, special collections, and other cultural heritage organizations. The last fifteen years have seen a wide range of initiatives among both small and large organizations to expose unique artifacts, including manuscripts, still photographs, film, audio and print. The practice of digitizing primary sources has come to be known in the vernacular of the library profession as "digital collections."

Staff at Duke University Libraries have collaborated during the last part of 2010 on re-imagining, re-engineering and re-implementing the web application by which the library provides discovery and access for its digital collections. Our work plan targets January of 2011 as a release date for the remade interface. As part of an outreach effort to the research and education communities that comprise the target audience for this application, we propose a poster and demo presentation at Digital Humanities '11.

For the development team at Duke, the project has presented an opportunity to think in depth about the landscape in which we publish our digital collections. The application that we seek to replace went live in January 2008, after an in-house development process that was expedited to head off the decommissioning of hardware. Since that time, we have considered how a more measured development process might enable us to upgrade the user experience.

We went through a lengthy process of gathering feedback from users and stakeholders of the existing digital collections site. We analyzed other libraries' digital collections efforts, content-focused sites like Flickr and Youtube, and retail sites like Netflix, Amazon and Zappos. We also gave considerable thought to the ways that open API's, social networking and concepts such as linked data transform the ways that students, instructors and researchers experience online resources.

Project manager Sean Aery has written extensively about the project and included many screenshots of the new page design on the library's Digital Collections Blog. His postings outline the research-heavy approach that we took in developing the interface design. In addition, we have worked to incorporate linked data concepts into the underlying representation of collections metadata. One of the objectives for this project is to develop an open API for the collections, to foster the development of applications by interested users. Presenting the framework to digital humanities scholars will prove invaluable in meeting this last objective. [See the http://library.duke.edu/blogs/digital-collections/category/website-redesign/].

One of the key decisions in our process was the choice to decouple the digital collections interface from other elements of our digital collections infrastructure. The application is a standalone framework developed in the Django platform, with Solr to provide the faceted searching functionality, and a mechanism for synchronizing data from other components. The decoupling of this application from tools for developing and managing content allows the development team to work in short, iterative cycles, free from dependencies on other aspects of the program's technology infrastructure.

- A significant selection of advertising-related materials from one of Duke's collection centers, the John W. Hartman Center for Sales, Advertising and Marketing History.
- Important documentary photography collections such as the Sidney D. Gamble Photographs of early-twentieth-century China.
- Many rare books not available in other venues.

Finally, and critically, all of these processes, decisions and source code updates have come in the service of a great body of important and compelling content for humanities researchers. Much of the material is housed in Duke's Rare Book, Manuscript and Special Collections Library, one of the world's leading libraries of its kind. Digitized collections include:

Duke has a very active Digital Production Center, and we plan to publish a wide variety of new and compelling materials in the first quarter of the new year. All of the content is freely available for educators, students and researchers.

Our process of learning from our users and patrons does not end with the release of the new platform.

We want to learn about ways we can improve the quality of the collections and the experience. Our goal is to make the digital collections a premier resource for research and learning in the humanities and social sciences. Digital humanities scholars comprise one of our important target audiences, and the DH11 conference provides us an important opportunity to engage in the discussion around this field of study.

Semantically Rich Tools for Text Exploration: TEI and SEASR

Ashton, Andrew Thomas
Andrew_Ashton@brown.edu
Brown University Library

Much of the existing work of researchers using the Text Encoding Initiative (TEI) guidelines has been made available on the web via processes such as XSLT transformations, which are intended to reproduce a work as an enhanced digital surrogate using HTML, PDF, or another publication format. Recent innovations in scholarly software design offer opportunities to exploit the semantic depth of TEI collections by creating new tools for textual analysis; tools that are designed not specifically for digital publication, but for creating mash-ups, data sets, and expressions of semantic data intended for machine-readability, rather than online readership. To explore these opportunities, the Brown University Library's Center for Digital Scholarship (STG) and the Brown University Women Writers Project (WWP) - with the support of the National Endowment for the Humanities' Digital Humanities Start-Up Grant program - are developing a prototype suite of software tools to explore TEI encoded texts in the Software Environment for the Advancement of Scholarly Research (SEASR). A demonstration of these tools, information about their availability, and a discussion about their usefulness to digital humanities projects, offers an opportunity to examine this approach to exploring digital texts.

2. Exploring Data Through Experimentation and Play

SEASR is a software environment for creating and sharing text and data mining tools. Textual analyses and visualizations created in SEASR can be shared on the web and adapted by other scholars to support their own research. The characteristic of SEASR that permits such exploration and interchange is its modular approach to software architecture. Modularity – the ability to share and reuse discrete pieces of software within a broader framework – enables relatively simple software tools to be combined to produce results that were previously the domain of complex and often proprietary software. Within the SEASR environment, analysis is modeled as a "flow": a pipeline, or a series of "components" arranged to produce a specific analysis (see figure 1). Each component in a flow is

a small piece of software that ingests and processes data, then passes it to the next component in the flow. Each component performs a single function, but through their recombination and sequencing SEASR enables innumerable potential analyses. Once a flow of components is constructed, it can be made publicly available as a web service and reused on other data. The SEASR framework permits components and flows to be integrated into the publication interface of a digital project (as a way of working with project data), or to be used as part of an independent web-based workbench through which a scholar could examine data drawn from multiple projects. In addition, scholars wishing to demonstrate a new analysis or method can create interactive web applications to make their work accessible and reproducible. And because flows can be flexibly combined and altered, they permit much greater latitude in scholarly inquiry, instead of channeling text analysis into predetermined sequences.

Designing a SEASR "flow"

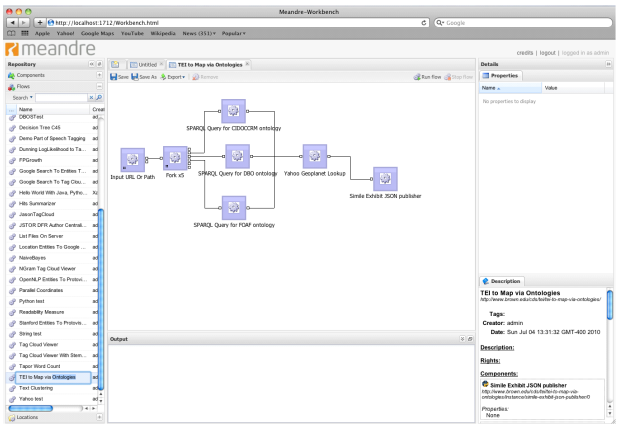


Figure. Designing a SEASR "flow"

3. Working with TEI

TEI's markup logic makes the semantic information encoded in a text tractable for the purposes of extraction, analysis, and reuse. When used in tandem with other data mining and visualization tools available in the SEASR environment (e.g., MONK, Simile, OpenNLP), these TEI components for SEASR offer new venues for exploring questions germane to literary and textual scholars.

As an example, a research scenario involving the WWP texts might include examining the language of specific genres and structures, possibly in relation to other information axes (e.g. time, geography). For example, how does the language of lyric poetry change during the course of the long 18th century? How do women situate themselves as authors (in dedications, letters to the reader, and acknowledgements)? To what extent do women make use of cultural referents from European history in situating their work, and how does this change over time? In pursuing these questions a researcher might use SEASR (either within a project interface such as Women Writers Online, or within a separate workbench application) to:

1. Extract and separate a collection of texts by genre. In addition to offering text-level genre information (based on a basic differentiation between prose, verse, and drama, which MONK also exploits), like many TEI projects the WWP marks genre-specific structures within the text such as poems, dramatic speeches, letters, recipes, essays). SEASR modules can thus identify and extract single genres of interest (all lyric poetry, for instance) or divide the entire corpus by genre (verse, drama, prose, etc.) at a specified level of granularity.
2. Extract specific pieces of text (quotes, verse stanzas, dedications) for downstream analysis.
3. Distill from the selected texts or text pieces the personal names, and separate these by type (references to historical figures, mythological figures, biblical figures; place names; etc.)
4. Sort a subset of data chronologically.
5. Pass the data through a component that adds morphosyntactic information to each word.
6. Generate a visualization such as a stacked area chart for each genre that analyzes changes in the association of certain adjectives with personal names, differentiated by gender.

4. Conclusion

The primary outcome of this exploration is a set of SEASR components designed to extract and process many commonly encoded semantic features using the TEI P5 Guidelines. This poster session will include a demonstration of the TEI tools for SEASR, as well as a discussion of the challenges of developing broadly applicable scholarly software tools such as these. Specifically, it will examine the difficulties and opportunities in designing software to work within a broad and diverse community of practice, such as the TEI community. The software, along with a set of sample analyses, will be available for download in a publicly available repository. In addition, a white paper, including an analysis of likely use-cases for these tools, will be available at the conclusion of the grant.

Extending the Life of the Broadside Ballad: The English Broadside Ballad Archive from Microfilm to Color Photography

Becker, Charlotte
becker.charlotte@gmail.com
University of California, Santa Barbara

Meyer, Shannon
meyer.shannon@gmail.com
University of California, Santa Barbara

The English Broadside Ballad Archive (EBBA) is a digital archive housed in the English Department of the University of California, Santa Barbara (http://ebba.english.ucsb.edu), under the direction of Professor Patricia Fumerton, and with funding from three NEH Reference Materials Grants (2006-8, 2008-10, and 2010-12). English broadside ballads were mass printed from the sixteenth through the ninenteenth century; in their "heyday" of the seventeenth century, they were multimedia productions that included woodcut images, a poem, and the title of a popular tune, to which the poem could be sung. The goal of the EBBA project is to recreate for users the lively interaction with these multimedia artifacts that approximates how early modern people might have encountered them at the height of their popularity. The EBBA archive is especially important because the surviving artifacts are extremely fragile and in most cases inaccessible to scholars, let alone the general public. The EBBA website provides facsimile images of the broadsides as they would have appeared when printed, transcriptions of the poem, facsimile transcriptions displaying the poem in the context of the original broadside but rendered in modern type, and audio recordings of the ballads being sung. Thanks to color images of ballad collections that EBBA has acquired during the past year, the EBBA team has begun to add another layer to the archive: now, in addition to reflecting the early modern experience with the ballads, EBBA can display the ballads as they appear in collections today.

We propose a multimedia poster presentation demonstrating the long-standing features of EBBA, and highlighting our recent acquisition of color images of the Roxburghe ballad collection through a contract with the British Library. At the British Library, the albums that comprise the Roxburghe collection are not made available to the public or even to most scholars because of their fragile condition; however, the texts in the collection remain highly important to scholars because they offer particular insight into early modern popular culture and demonstrate developments in print culture over the three centuries during which the collected ballads were printed and circulated. EBBA's acquisition, manipulation, and mounting of the Roxburghe color images is exciting and challenging from both the user and developer points of view, and will be the focus of our presentation.

For scholarly users of the site, these photographs are important because they show the results of the collection process. In the case of the Roxburghe collection, this process often involved cutting broadsheets in half and pasting the halves into albums, with the two sides of the ballad either on facing pages or one side above the other on a single page. Sometimes ballads were pasted sideways into the album, uncut and folded at the edges to fit inside the closed album. EBBA has previously, for the collection of Pepys ballads, offered microfilm-based black and white images (called "ballad facsimiles" on our site) to simulate the appearance of the original whole broadside. But adding color images and "album facsimiles," as in the case of the Roxburghe collection, helps EBBA give users a better sense of the long life and changing cultural contexts of broadside ballads. One aspect of this long life is the handwritten emendations made by collectors and overzealous antiquarians. Whether commentary or additions that assist in reading the ballad, these emendations are an essential part of understanding the reception and use of the artifacts over time. EBBA's text transcription rules are geared toward providing a searchable, readable text of each ballad as it was originally printed, yet the EBBA team wanted to ensure that the prolific and often illuminating handwriting in the Roxburghe collection could be accessed by users. The color images capture the physical object in great detail, rendering the handwriting entirely visible and accessible to scholars wishing to read it; thus, users can now see this important feature of the Roxburge ballads for themselves.

As our presentation will show, these color images have been a unique opportunity for the EBBA team to think about our responsibility to present these artifacts in a way that balances scholarly utility, visual fidelity, and the EBBA site's technical capability. The British Library's "Turning the Pages" software and Virtual Books webpage is a well-known example of the kind of virtual access to textual artifacts that we aim to

provide; however, the nature of the Roxburghe albums makes such a presentation particularly challenging. We are dealing with albums where the ballads were arranged in unpredictable ways when collected, where page discoloration makes aesthetic uniformity difficult to achieve, and where parts of the ballads are often obscured or distorted by folding or insertion in the albums' gutter. In addition to finding ways to account for these visual aberrations, the EBBA team spent a great deal of time deliberating over optimal file size for image delivery and attempting to find the most advantageous balance between server speed and image detail. These practices and decisions are crucial ones as EBBA continues to acquire color images of other ballads, including the Euing collection from the University of Glasgow, and the Britwell ballads from the Huntington Library.

Given the irregular and sometimes haphazard nature of ballad collection, decisions such as these will continue to face the EBBA team as the project moves forward with these and hopefully other extant collections in the future. We are excited to share EBBA's progress, as well as our dynamic process of decision making, as we continue to make early modern broadside ballads available to the public.

Virtual Touch. Towards an Interdisciplinary Research Agenda for the Arts and Humanities

Bentkowska-Kafel, Anna
anna.bentkowska@kcl.ac.uk
Centre for Computing in the Humanities, King's College, London, UK

Giachritsis, Christos
c.giachritsis@bham.ac.uk
SyMoN lab, University of Birmingham, UK

Prytherch, David
david.prytherch@bcu.ac.uk
User-lab, Birmingham Institute of Art and Design, Birmingham City University, UK

The term 'haptics' encompasses two areas of study: human and machine haptics. The first relates to the study of the perception of the world through the sense of touch. It includes proprioception (one's awareness of one's own body position in space) as well as cutaneous information (one's awareness of skin deformations). Machine haptics relates to the design and development of devices that simulate the haptic properties of physical objects. In principle, they are incorporated in virtual environments and allow users to experience tactile properties of virtual objects such as size, shape, weight, compliance and texture.

2. Virtual Artefact. A Different Approach

The virtual artefact has firmly established itself as a research tool within several disciplines of the Arts and Humanities. Many art and material culture historians and professionals rely on digital records and visualisations of artefacts in their research, teaching and practice. We have witnessed, from the 1990s onwards, how the virtual artefact has increasingly become photo-realistic and interactive, and how it continues to evolve. The virtual artefact can now be part of a complex, collaborative research environment. With the enhanced technical specifications comes the interest in exploring the research potential of virtual artefacts further. We are here concerned with enhanced simulation of the real experience of physical objects through the application of haptic interfaces, or virtual touch technologies. We believe that the addition of virtual touch would also contribute to greater

usability of the existing, often neglected electronic resources and libraries of 3D artefacts.

Research into the use of haptic interfaces–that is devices engaging the sense of touch in virtual environments–in the Arts and Humanities is in its infancy. Although virtual simulation of physical touch has resulted in important advances in other disciplines–such as medicine, neuro-science, telemanipulation control systems and product design–the potential of such applications to humanities scholarship has not yet been explored. Very few researchers in the Arts and Humanities have had an opportunity to experience haptic devices first hand and to develop a critical understanding of such systems and the perceptual processes involved. There have been some important and promising developments in the area of heritage science, such as the Haptic Museum in the US, being the work of Margaret McLaughlin *et al*. (2000) at the University of Southern California, Annenberg and the Los Angeles County Museum of Natural History; the Museum of Pure Form (Bergamasco *et al*., 2005) and 'Touching the Untouchable: Increasing Access to Archaeological Artefacts by Virtual Handling' in the UK, supported by the Arts and Humanities Research Council and Engineering and Physical Sciences Research Council.

The authors are investigating certain fundamental questions:

- Is virtual touch likely to enhance the ways in which we carry out and communicate research in those areas of the Arts and Humanities that employ 3D visualisations of material culture?
- How much do we learn about artefacts by touching them?
- Would simulating this experience through a haptic interface enhance virtual fieldwork?
- Who can possibly benefit from this experience and how?
- If 'seeing with vision that feels, feeling with fingers that see' (Goethe, 1788) is possible, to what extent can this experience be mediated by a haptic system?

3. To Touch or not to Touch?

A notice 'DO NOT TOUCH' is familiar in many museums and heritage sites and the reasons why this measure is necessary are commonly understood. Most museum managers and curators embrace the notion that touching object collections are important. However, many of the most tactually alluring artefacts present conservation problems that are difficult to overcome. Yet, the existence of 'DO NOT TOUCH' simply demonstrates the visitor's compelling 'NEED TO TOUCH' the artefact. This need originates from the fact that the sense of touch can provide us with haptic-specific information about an object such as absolute size, material compliance, texture, temperature and weight, which cannot be reliably obtained through the audio-visual senses. Therefore, a solution which would allow visitors, curators and researchers in Arts & Humanities to experience these tactile properties without compromising the integrity of the artefact is likely to be very welcome and highly valuable.

The latest haptic technologies can provide a way forward which satisfies both the visitors' and researcher's desire to handle and the museums' necessity to conserve collections for future generations. An early 'proof of concept' pilot study was conducted with randomly arriving museum visitors, which looked at the potential for the substitution of real object handling with touching virtual 3D replicas via a low resolution, low cost, commercially available haptic interface (PhanTom Omni) coupled with a stereoscopic (3D) visual display (Fig. 1).

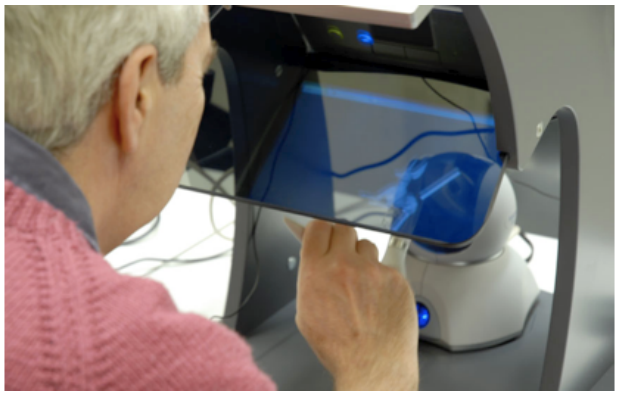


Fig. 1 'Feeling' a virtual replica of an object from the Potteries Museum, Stoke-on-Trent, UK, using the portable haptic system based on a PhanTom Omni, SenseGraphics 3D Mobile Immersive Workbench. Photo: D. Prytherch

The visitor feedback offered interesting insights into the potential of the haptic display of virtual objects to enhance the experience of museum visitors by allowing them to interact with virtual artefacts through touch. An informal trial with a blind member of museum staff demonstrated that he was able to identify the shape of the object and gain some insight into the surface carving and texture of the object, despite the real artefact being untouchable. The same portable haptic system was used in the classroom when teaching a masters module in Digital Arts and Culture and received valuable feedback from the students (Fig. 2).

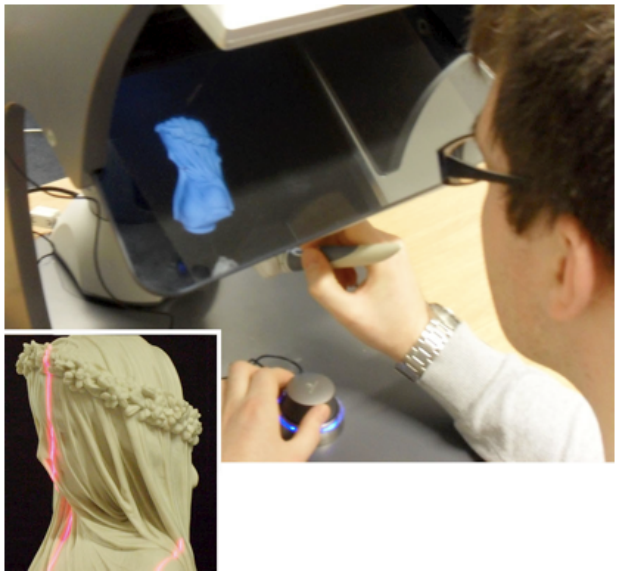


Fig. 2 Haptics in the classroom. A postgraduate course in Digital Arts and Culture, Centre for Computing in the Humanities, King's College London, UK, 24 November 2010. Photo: A. Bentkowska-Kafel. Insert: Laser scanning of the Bride (Potteries Museum, Stoke-on-Trent, UK) to create a virtual 3D replica. Photo: D. Prytherch.

Archaeologists, art historians, restorers, palaeographers and other specialists who examine artefacts through touch to assess and authenticate their material, execution and other tactile qualities, might also be interested in haptic access to material ex situ when direct access is not possible. The capacity of human haptic perception–the ability to perceive the world through the sense of touch–to differentiate qualities of material, is incredibly rich and we learn much about the form and composition of an artefact from our sensitivity to thermal conduction and fine surface textural qualities through touch. Current commercial haptic devices do not make effective use of tactile (or cutaneous) cues, that is the physical properties of objects which are perceived through skin mechanoreceptors. Technologies that do operate at this level are either experimental, lab-based systems or, if commercially available, are extremely expensive, as well as being complex to setup and maintain. Certain limitations inherent in the whole concept of virtual representations of objects, notably weight and balance, which are dependant not on the virtual object itself, but on the properties of the specific device with which we feel it, need to be resolved.

The virtual artefact has opened up unprecedented possibilities for new research into the material culture of the past. Digital technology supporting visualisation of heritage is becoming ever more sophisticated and many projects of this kind seek to better the accompanying scholarly apparatus. The publication of the *London Charter for the Computer-based Visualisation of Cultural Heritage* (version 2.1, February 2009), is a step in this direction. Some scholars working in this area claim that their visualisations of ancient sites re-create the real human experience of 'being there'. Some believe that digital visualisation itself may 'include sight, hearing, and potentially in the future, smell, taste, and touch' (Mudge, 2011). The lack of tactile experience is one of the significant issues inherent to digital visualisation of heritage. In addition, our limited knowledge of the real-life processes that we are trying to simulate through the use of digital media is always a challenge. We need to better understand the perceived discrepancies between the real object and its virtual record.

4. 'Feeling with a Seeing Hand' (Goethe, 1788)

In the process of gaining knowledge about the *real* world, *vision* allows us to explore the environment instantly and provides us with information about object properties such as 3D shape, colour, texture, condition, relative size, relative distance as well as events such as motion or changes of all these visual properties. Even when looking close-up is possible our knowledge is still incomplete. This is because vision alone does not allow us to perceive important physical properties of materials such as weight, centre of mass, temperature, surface roughness, hardness and actual size. In order to perceive these properties we need to engage *touch*. Touch allows us to perceive the properties which are essential for effective *interaction* with the physical world. By exploring an object through touch, we can form a better understanding of its purpose and use, and develop insights to inform us about its past users. For example, it is not possible to appreciate the physical endurance of an ancient warrior just by viewing a shield he might have used. For this, it is necessary to engage touch: lift and carry it.

Although still relatively primitive (comparing to advances in visual displays), current haptic interfaces (HI) allow a range of haptic and tactile interaction with virtual objects, from the exploration with one or multiple fingers with one or two hands (Giachritsis *et al*., 2009; Giachritsis *et al*., 2010; Monroy *et al*., 2008) (Fig. 3) to more immersive experience involving movements from arm joints.

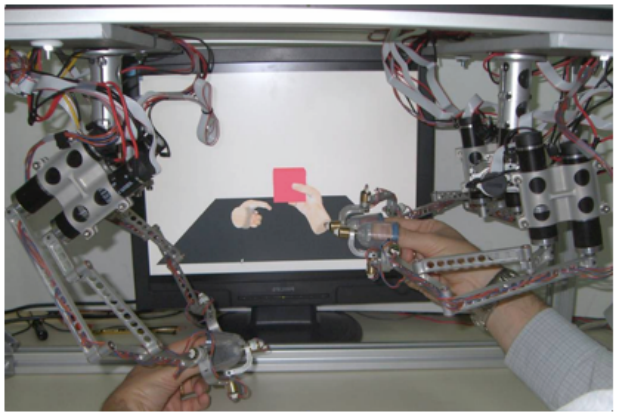


Fig. 3 Bimanual manipulation of a virtual object using precision grip with the haptic interface device Master Finger 2.

Haptic applications vary from simple touch of the virtual object to real-life probes used as medical or artistic tools for working on object surface. The extensive industrial and academic research in advancing HI shows the impact of these technologies on the way people learn, train, work and entertain themselves in virtual environments. Despite their limitations, HI technologies already offer significant advantages to users of virtual environments by allowing them a more realistic interaction with virtual objects.

Recent advances in recording art objects, made possible by such technologies as photogrammetry, touch probing and 3D structured light or colour laser scanning, have augmented the already considerable body of virtual artefacts, created primarily through Virtual Reality (VRML) modelling, with records of unprecedented accuracy. The E-Curator project carried out by the University College London Museum Collections and partners, used 3D colour scans of various artefacts captured with an Arius scanner, to develop a traceable, grid-based dissemination and visualisation system. This system enables museum curators and conservators to identify and assess objects remotely in a collaborative, networked environment. The addition of haptic interfaces engaging the sense of touch seems an obvious future development. The question is which particular aspects of human touch these devices should simulate so that they could be useful to heritage professionals?

5. Towards a Research Agenda for the Arts and Humanities

If the promise of more intuitive and multisensory computing–than the one with which we are familiar today–is real, we should investigate how these future developments in information technologies may affect our own disciplines. We argue for the need to explore the potential benefit of haptic interaction as an addition to audio-visual experiences of virtual artefacts. How much do we learn about artefacts by touching them? Would simulating this experience through a haptic interface enhance virtual fieldwork? Who can possibly benefit from this experience and how? If Goethe's (1788) ‘seeing with vision that feels, feeling with fingers that see’ is possible, to what extent can this experience be mediated by a haptic system? We will not be certain of the answers, nor have influence over the future of multisensory computing, unless we explore whether current developments and wide-ranging research meets our varied expectations.

The authors propose a research agenda in this area is made from the four positions: 1) digital scholarship of material culture; 2) the contribution of human vision and touch in the perception and appreciation of real and virtual environments; 3) haptic interface design and applications, and 4) haptic access to 3D records of art objects within networked collaborative environments.

The authors seek to investigate the level of interest of those engaged in historical and cultural studies in researching the potential benefits of adding virtual touch to virtual artefacts for the advancement of cultural heritage scholarship and education. Expressions of such interest and comments are welcome.

References

Brewster, S.A. (2005). 'The Impact of Haptic ‘Touching’ Technology on Cultural Applications'. *Digital Applications for Cultural Heritage Institutions.* Hemsley, J., Cappellini, V., and Stanke, G. (ed.). Aldershot: Ashgate, pp. 273-284.

E-Curator Project, University College London. `http://www.museums.ucl.ac.uk/research/ecurator/` (accessed 14 March 2011).

Frisoli, A., Jansson, G., Bergamasco, M. and Loscos, C. (2005). 'Evaluation of the Museum of Pure Form Displays Used for Exploration of Works of Art at Museums'. *Proc. World Haptics Conference 2005.* Pisa, 18-20 March 2005. `http://ima.udg.edu/~closcos/Publications/Frisoli-A-Evaluation-PureForm_regular.pdf`.

Giachritsis, C., Barrio, J., Ferre, M., Wing, A., and Ortego, J. (2009). 'Evaluation of Weight Perception During Unimanual and Bimanual Manipulation of Virtual Objects'. *Third Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual*

Environment and Teleoperator Systems. Salt Lake City, UT, 18-20 March 2009.

Giachritsis, C.D., Garcia-Robledo, P. Jr., Barrio, J., Wing, A.M. and Ferre, M. (2010). 'Unimanual, Bimanual and Bilateral Weight Perception of Virtual Objects in the Master Finger 2 Environment'. *19th IEEE International Symposium on Robot and Human Interactive Communication.* Principe di Piemonte - Viareggio, Italy, 12-15 September 2010.

Goethe, J.W. von (1988). *Roman Elegies, VII [1788].* London: Routledge.

London Charter for the Computer-based Visualisation of Cultural Heritage. http://www.londoncharter.org (accessed 14 March 2011).

McLAUGHLIN, M., L., Sukhatme, G., Shahabi, C., Medioni, G. and Jaskowiak, J. (2000). 'The Haptic Museum'. *Proc. EVA Conference on Electronic Imaging and the Visual Arts.* Florence, 2000.

Mudge, M. (forthcoming). 'Transparency for Empirical Data'. *Paradata and Transparency in Heritage Visualisation.* A. Bentkowska-Kafel et al. eds (ed.). Aldershot: Ashgate. http://visualizationparadata.wordpress.com .

Monroy, M., Oyarzabal, M., Ferre, M., Campos, A. and Barrio, J. (2008). 'MasterFinger: Multi-finger Haptic Interface for Collaborative Environments'. *Proceedings of Eurohaptics '08.* Madrid, 10-13 June 2008 , pp. 411-419.

Prytherch, D. and Jefsioutine, M. (2007). 'Touching Ghosts: Haptic technologies in museums'. *The Power of Touch: Handling Objects in Museum and Heritage Contexts.* Pye, E. (ed.). Walnut Creek, CA: Left Coast Press, pp. 223-40.

Touching the Untouchable: Increasing Access to Archaeological Artefacts by Virtual Handling. http://www.heritagescience.ac.uk/index.php?section=97 (accessed 14 March 2011).

Improving the AAC-FACKEL, a Scholarly Digital Edition of the Satirical Journal "Die Fackel"

Biber, Hanno
hanno.biber@oeaw.ac.at
Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences,Vienna, Austria

In the following a presentation of the latest developments to improve an existing scholarly digital edition will be given. The scholarly edition in question is the AAC-FACKEL, the digital edition of the historical literary journal "Die Fackel". The presentation will be devided into three parts representing three consistent steps in the development. First, the general principles and the specific edition and design considerations concerning the online publication of the AAC-FACKEL will be presented. Second, the particular questions of editing and exploring this important and interesting source of literary history of the German language by means of a sophisticated research tool will be addressed. Third, a plan and the considerations for improvement of this successful and widely used online edition, which is based upon the principles of corpus research and text technology, will be presented. The digital edition of the historical literary journal "Die Fackel" ("The Torch") has been developed in a collaboration of researchers, programmers and designers within the framework of the AAC-Austrian Academy Corpus which is operated by the Institute for Corpus Linguistics and Text Technology at the Austrian Academy of Sciences in Vienna. "Die Fackel" was originally published and almost entirely written by the satirist and language critic Karl Kraus in Vienna from 1899 until 1936. The AAC-FACKEL is online since January 2007 and offers free online access to its 37 volumes, 415 issues, 922 numbers, comprising more than 22.500 pages and 6 million tokens. The digital edition contains a fully searchable database of the journal with various indexes, search tools and navigation aids in an innovative and functional graphic design interface, where all pages of the original are available as digital texts and as facsimile images, which is one important principle of the AAC's resources and publication initiatives. The work of Karl Kraus can be regarded as one of the most important contributions to world literature. It is a source for the history of the time, for its language and its moral transgressions. Karl Kraus covers in his typical and idiosyncratic style in

thousands of texts the themes of journalism and war, of politics and corruption, of literature and lying. The interface provides the scholar with a complex research environment to read, study and access the texts from various entry points. The journal comprises a great variety of essays, notes, commentaries, aphorisms, poems and other forms. The scholarly digital edition allows new ways of philological research and analysis. The presentation of this resource, where the question of how a valuable historical text source is made visible and alive in a digital media environment is addressed, gives insights into the features of the edition, where the potential of lexicographic word searches within the various texts of the edition as well as basic and more advanced elements of the corpus research approach followed by the AAC is demonstrated. The research tool offers access to this resource by means of a multifunctional interface. The interface has five individual frames synchronized within one single window. The frames can be opened and closed as required. The 'Paratext' section situated within the first frame provides additional information about the background of the edition and scholarly essays about the journal. The 'Search|Index' section gives access to a variety of indexes, databases and full-text search mechanisms. The results of these queries and lists are displayed in the adjacent 'Results' section. A sophisticated 'Contents' section has been developed in order to show the reader the whole range of the journal ready to be explored and give access to the whole run of issues and all of the contents of the journal in chronological order. The 'text' section has a complex and powerful navigational bar at the top so that the reader can easily navigate and read within the journal either in text-mode or in image-mode from page to page, from text to text (soon to be implemented), from issue to issue, and with the help of hyperlinks. The text can be searched in the following four modes: in full text search mode, by means of a list of word forms and of an inverted list of word forms. The functionality of the indexes and the technical specifications of the edition are described in the 'Paratext' section on the technical details. The following indexes are in preparation: an index of titles including all texts of the entire journal; an index of errors that were corrected by Karl Kraus in "Die Fackel"; an index of errors that were not corrected by Karl Kraus but found and provided with suggested readings by the editors; an index of variants of Fackel issues, including the later editions, all confiscated issues and censored passages as well as the various regional editions of "Die Fackel"; an index of inserts including all regular inserts that are considered as part of the text of "Die Fackel" together with other inserts, which were added on occasion, as well as additions; an index of illustrations; an index of special issues; and an index of extra editions, including separately reprinted articles from "Die Fackel"; finally, an extensive and complete index of names mentioned in "Die Fackel", together with philologically sampled additional data, is in preparation and will consist of personal names, based upon the material collected by Franz Ögg, but corrected and largely extended, as well as indexes of fictitious names, institutions, periodicals, and works of literature, works of art and so on. The contents list of this edition is based on the original text titles given by Karl Kraus in "Die Fackel" and in the separate contents pages of the quarterly volumes of the journal, which were collected by the editors of the AAC-FACKEL. The corpus based research possibilities offered by this digital edition and its underlying principles will be dealt with in this presentation. The corpus research unit at the Institute for Corpus Linguistics and Text Technology of the Austrian Academy of Sciences is concerned with establishing and exploring large electronic text corpora and with conducting scholarly research in the field of corpora and digital editions. Among the sources collected by the AAC, which systematically covers various domains, genres and types, are more than 500 million running words of newspapers, literary journals, novels, dramas, poems, advertisements, essays on various subjects, travel literature, cookbooks, pamphlets, political speeches as well as a variety of scientific, legal, and religious texts, to name just a few forms. The specific principles of the digital editions are determined by the conviction that the methods of corpus research lead to valuable resources for scholars. The AAC has developed model editions to meet these aims, which provide well structured and well designed access to the sources, and will continue to do so by improving existing editions, thereby contributing to the development of digital resources for research into language and literature.

Constructing DARIAH—the e-Infrastructure for the Arts and Humanities

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

Fritze, Christiane
fritze@sub.uni-goettingen.de
State and University Library Goettingen

Romary, Laurent
laurent.romary@inria.fr
INRIA

The poster will elicit our vision for the DARIAH infrastructure and the first steps towards its implementation. DARIAH (Digital Research Infrastructure for the Arts and Humanities; http://www.dariah.eu) is a European project funded under the ESFRI programme (http://cordis.europa.eu/esfri/), which aims to design a virtual bridge between various humanities and arts resources across Europe. DARIAH is currently in its transition from the preparatory phase to the construction phase, which will be completed by the establishment of the legal framework DARIAH ERIC by end of 2011.

Just like astronomers now require a virtual observatory to study the stars and other distant objects in the galaxy on the basis of a wide variety of existing observations, researchers in the arts and humanities need a digital infrastructure to bring together and collaboratively work with dispersed scholarly resources (e.g. digital content, services, methodologies). DARIAH will be such an infrastructure with a European dimension to support research practitioners at all levels, from beginners through to those employing advanced techniques and methodologies. The grand vision for DARIAH is to facilitate long-term access to, and re-use of, all European arts and humanities digital research data and primary sources.

A typical use case is elicited below (the poster will present several of such scenarios taken from actual e-humanists Europe-wide)

2. The Digital Postgraduate

Daniela is a postgraduate researcher (PhD candidate) in material culture at a Greek university. She holds a first degree in classical archaeology and a Master's in anthropology. Daniela uses ICT tools very efficiently and considers them vital for her research. She is currently researching material culture and its relationship to the perception of space and landscape in an area of northern Greece. Her topic lies within the areas of archaeology and cultural anthropology; therefore, Daniela's sources include artefacts, interviews with local people, and extensive visual material. As her work is largely interdisciplinary in nature, there is a propensity for making fortuitous discoveries while looking for something unrelated; serendipity is very important in her work. This approach means that she needs to login to different services simultaneously—even login multiple times a day if disconnected—and she must keep multiple open windows in her browser. Furthermore, she must be able to evaluate the authenticity and value of any item she discovers. Thanks to the technical environment developed in the VCCe-Infrastructure of DARIAH, Daniela benefits from both a single sign-on environment for accessing all the necessary digital assets,as well as a virtual portfolio where she can gather all selected sources. The metadata and provenance data associated with an item allows her to evaluate the authenticity and value of an asset, and to link to related items (e.g. other sources for the same asset, other formats, research addressing that asset) once she has found something of interest. From this portfolio, she can publish geographical views on her data, which she exchanges and discusses with other colleagues in Europe.

The mission of DARIAH is to enhance and support digitally-enabled research across the humanities and arts. DARIAH aims to develop and maintain an infrastructure in support of ICT-based research practices and is working with communities of practice to:

1. Explore and apply ICT-based methods and tools to enable new research questions to be asked and old questions to be posed in new ways;
2. Improve research opportunities and outcomes through linking distributed digital source materials of many kinds
3. Exchange knowledge, expertise, methodologies and practices across domains and disciplines.

DARIAH is also not singly discipline-focused; instead DARIAH seeks to support all disciplines across the humanities, encouraging interdisciplinarity and the exploration and sharing of content, tools and methods. Research practice in the arts and humanities is about criticism and meaning, interpretation and re-

interpretation, and about extracting meaning from often incomplete and fuzzy data. It requires researchers to seek out a wide range of primary and secondary sources, to organise and structure these, to analyse and interpret them, and to publish the results. In this era of pervasive broadband connectivity, the way in which these processes are undertaken is changing, and in some cases, the processes themselves are changing. Increasingly, research practitioners are using the power of the internet, new tools, and the range of digital information that is available to them to create their own personal network spaces, to digitally publish highly interactive, multimedia-themed collections (critical editions) of research information and knowledge, and to visualise and reconceptualise their interpretations and analysis. New forms of collaboration are also emerging as the tools available encourage and enable 'web-working' across the globe.

Hence the key strategic aim of DARIAH is to support researchers in the creation and use of research data and tools, and to apply and use ICT-enabled methods to analyse and interpret digital source materials.

DARIAH will be an infrastructure to promote, support, and advance research in the digital humanities. Digital humanities is a long-established research field, with its origins in the Forties of the last century. Over the past 60 years it has progressed and a large variety of digital humanities centres and related organizations have developed. However, we do not perceive the digital humanities to be a closed field of existing centres but rather an open and developing research environment. Everybody interested in using digital means for arts and humanities research is part of the DARIAH community of practice. In this view, the DARIAH infrastructure would be a connected network of people, information, tools and methodologies for investigating, exploring and supporting work across the broad spectrum of the digital humanities.

The DARIAH network will be designed to be as a decentralised network of competency centres (VCC – Virtual Competency Centres), which will allow services to stay close to end-users (researchers) either geographically or thematically. Common technologies (e.g. for authentication or federation of archive contents) and good practices (standardised formats, digital assets management workflows) will ensure coherence across the support services offered by the competency centres.

When DARIAH is operational after the construction phase, technical products by DARIAH will be manifold:

- technological services and tutorials that help existing humanities data archives to link their systems into the DARIAH network;
- a package of software and consultancy/training, which supports emerging data centres in establishing their own technology environment quickly;
- an interoperability layer that will connect data centres;
- means of linking into DARIAH for those countries/ disciplines that do not yet have e-humanities infrastructure and cannot afford it in the near future;
- best practices and guidelines for individual researchers that foster data interoperability and preservation across the DARIAH network;
- a network of expertise linking each scholar to an active and vibrant community of international digital humanists.

DARIAH will make an important contribution towards e-humanities, providing additional services to analyse, annotate and share arts and humanities research activities. DARIAH will stimulate and provide expertise on all aspects of e-humanities, from best practices for digitisation to metadata standards and advice on analysis methods and systems.

The Arcane Gallery of Gadgetry: A Design Case Study of an Alternate Reality Game

Bonsignore, Beth
elizabeth.bonsignore@gmail.com
College of Information Studies, UMD

Goodlander, Georgina
GoodlanderG@si.edu
Smithsonian American Art Museum

Hansen, Derek
shakmatt@gmail.com
College of Information Studies, UMD

Johnson, Margeaux
geauxgeaux@gmail.com
University of Florida

Kraus, Kari
karimkraus@gmail.com
College of Information Studies, UMD

Visconti, Amanda
amandavisconti@gmail.com
Department of English, UMD

1. Introduction

This DH2011 paper describes and analyzes the design process and delivery for The Arcane Gallery of Gadgetry (AGOG), a “mini Alternate Reality Game,” that has become the seedbed for a larger ARG currently under development by the authors with a planned launch date of spring 2011. With generous two-year support from the National Science Foundation (NSF), the multi-disciplinary team is conducting exploratory, qualitative research at the University of Maryland on the use and creation of ARGs as participatory design spaces, information literacy systems, and vehicles for scaffolding student learning. The collaborators also include Georgina Goodlander, Interpretive Programs Manager at the Smithsonian American Art Museum; and Margeaux Johnson, Science and Technology Librarian at the University of Florida.

An Alternate Reality Game is a form of interactive storytelling whose narrative elements are distributed over multiple "real-world" platforms, including books, mobile devices, and networked computers. They often require participants to solve puzzles, answer riddles, and track down information in order to advance the storyline. Examples of popular commercial ARGs include games such as “The Beast,” “I Love Bees,” and “The Lost Experience.” Although ARGs have been used primarily for entertainment, they can also provide unique and powerful educational opportunities. World Without Oil and ARGOSI are among the growing number of ARGs created with educational, “serious game” goals in mind.

The case study of AGOG focuses on a distinct aspect of ARG design that is well positioned to benefit from theoretical insight and methodological inquiry: namely, embedding counterfactual story bits into a larger historical framework.

The challenge of counterfactual design is to decide how to purposefully, meaningfully, and responsibly depart from the historical record when developing ARGs within the context of libraries, schools, museums, and archives--cultural institutions that place a high value on trustworthiness and accuracy of information, including digital information. The slippage between fiction and reality that is the sine qua non of the genre should be just as much the result of premeditation as any other aspect of game design. To that end, we propose a theoretical and methodological framework for counterfactual design that draws on the neuroscience research of Ruth Byrne, the object-oriented philosophy of Ian Bogost, and the cooperative design techniques pioneered by Allison Druin in the context of Human Computer Interaction (HCI).

2. Design Case: The Arcane Gallery of Gadgetry (AGOG)

The design attributes of AGOG, a number of which have already been implemented and play-tested, can be conveyed in part with the help of scoping notes:

- Place and time period: Civil War and Post-War Reconstruction United States. While the Civil War (understood broadly to also include antebellum and post-war periods) helps anchor the ARG temporally, geographically, and—if only tangentially—thematically, these parameters are intended as guidelines, made flexible by the game’s liberal use of time slips, motivated anachronisms, and counter-factual scenarios.
- Thematic features: The game includes the following narrative motifs: historical figures communicating across space and time, messages and artifacts intended for posterity, and secret societies whose

members may act as stewards of these information-bearing artifacts.

- Activities, puzzles, and history hacks: AGOG incorporates player-created/-curated artifacts, scavenger-hunt like missions; information search and retrieval exercises; and cryptographic challenges.
- Aesthetic: The game is steeped in a nineteenth-century retro-futuristic, steampunk style.
- Learning objectives: A primary objective of AGOG is to use the machinery and conventions of ARGs to scaffold information and new media literacy instruction, as well as to teach subject knowledge in history, science, technology, and math.
- Tools and technologies: The extended version of AGOG will make use of the technological affordances of smartphones--such as camera, phone, GPS, texting, and web-browsing functionality--to enhance interactivity and integrate the offline and online worlds in creative ways.

The game's mythology is grounded in the history of the U. S. Patent Office. The Smithsonian American Art Museum, where Georgina Goodlander works, is housed in the Historic US Patent Office Building. During the time it functioned as a "Temple of Invention" (1836 to 1932), thousands of patents were submitted, along with miniature models of the designs, which were put on display. The building also variously served as a place of employment, curiosity, ministry, and sociability for a number of historically significant figures and personalities, including Abraham Lincoln, Walt Whitman, and Clara Barton. The stately halls and galleries of the Patent Office were transformed into makeshift barracks and hospital rooms during the Civil War, the grim realities of which were temporarily overshadowed by the repurposing of the space for Lincoln's second inaugural ball in 1865.

Collecting all of this historical data in a document, the design team began looking for events and places through which it could dig a fictional or counterfactual tunnel. In 1877, "the noblest of Washington buildings" was dealt a severe blow when a fire broke out and destroyed the collection of 12,000 rejected patent models in the attic and damaged another 114,000. It is the fire that provided us with the means to traverse fiction and reality, functioning as a joint in which we could embed a "rabbit-hole" that would draw players into the game: a mysterious document allegedly dating back to the fire of 1877, which cryptically refers to a "Cabinet of Curiosities." Out of the notional fragments of patent models and other mechanical remains that ostensibly survived the conflagration, players are asked to help reconstruct and curate the Arcane Gallery of Gadgetry by imagining what sorts of wondrous, retro-futuristic inventions might have populated it and then creating those artifacts using found objects (in the spirit of assemblage art) that have first been identified in a database of historic patents. Possible categories of inventions include but are not limited to communications devices, weapons and ammunition (think secret Civil War technologies), cryptographic devices, and medical equipment. Figure 1 shows the chronological and conceptual relationships among the factual and counterfactual narrative elements in AGOG. Other resources relevant to AGOG, including video footage of player-created artifacts for the gallery, can be accessed online at http://www.karikraus.com/?p=69.

Figure 1. Chronological and conceptual relationships among factual and counterfactual narrative elements in AGOG.

3. Counterfactual Design

The DH2011 paper triangulates between the neuroscience research of Ruth Byrne, the object-oriented ontology (OOO) of Ian Bogost, and the HCI cooperative design techniques developed by Allison Druin, distilling from them the following set of principles and practices for counterfactual design.

Cognitive Science: In The Rational Imagination: How People Create Alternatives to Reality, Ruth Byrne explores the "faultlines" or "joints" of reality, those "aspects of reality that are more readily changed in imaginative thought." Byrne's premise is that there are patterns or regularities in terms of where we locate those faultlines: some attributes of reality just seem inherently more "mutable" to us than others. When we partition, classify, and organize our world, for example, we readily invent new instances or members of a category. In the case of AGOG, the design team identified a real category—patent collections lost (or

presumed lost) in the fire of 1877—and created a new member for it: the Arcane Gallery of Gadgetry.

Object-Oriented Ontology: In his forthcoming book Alien Phenomenology, Ian Bogost offers a practice-led approach to philosophy, one that challenges the notion that the inevitable product of philosophical inquiry is a paper, article, or book ("Latour Litanizer"). Instead, Bogost describes and models an alternative approach, carpentry, which "refers primarily to the construction of artifacts that illustrate the perspectives [or inner lives] of objects" ("Latour Litanizer"). As he puts it elsewhere, "as critics, our job is to amplify the black noise of objects to make the resonant frequencies of the stuffs inside them hum in credibly satisfying ways" ("Interview"). Having co-authored Racing the Beam with Nick Montfort in 2009—an MIT platform studies book on the Atari Video Computer System--Bogost compellingly speculates on how carpentry might help us better apprehend the system's internal hardware design: "What's it like to be a [Television Interface Adapter]? Or a MOS Technologies 6502 microprocessor? How would one characterize such a thing? Would it even be possible?" ("Interview").

The DH2011 presentation describes how this "pragmatic" mode of OOO might be applied through circuit-bending—characterized by artists in a popular YouTube video as "parallel worlds within a circuit"--to the Arcane Gallery of Gadgetry.

Cooperative Design: Inspired by the prospect of what we might learn about ARGs and associated technologies by including children in the design process, we partnered with KidsTeam at the HCIL in summer and fall 2010. Adopting an exploratory question related to AGOG, we asked the children how they would collaborate with other in-world players, especially if they had certain in-game constraints, such as only being able to use found objects or materials from a certain time period. Preliminary findings suggest that the joints or faultlines of reality exist across multiple semiotic domains, and that the gap separating fiction and reality is often managed through recourse to metaphor. The DH2010 presentation will elaborate on these findings and their implications for the design of tools that support ARGs.

4. Funding

This work is supported by the National Science Foundation.

References

Nicola Whitton (2010). 'ARGOSI :: Alternate Reality Games for Orientation, Socialisation and Induction'. . . http://argosi.playthinklearn.net/index.htm.

Bogost, Ian. (16 December 2009). 'Latour Litanizer'. . http:// http://www.bogost.com/blog/latour_litanizer.shtml.

Byrne, Ruth M. J. (2007). *The Rational Imagination: How People Create Alternatives to Reality.* The MIT Press.

DrRek (10 March 2006). *What is Circuit Bending?.* YouTube. http://www.youtube.com/watch?v=w6Pbyg_kcE.

Gratton, Peter (26 April 2010). *Ian Bogost: The Interview. Philosophy in a Time of Error (Blog Post).* . http://philosophyinatimeoferror.wordpress.com/2010/04/26/ianbogost-the-interview/.

Ken Eklund (Game Designer, Creative Director and Producer) (2007. Web. 1 Nov. 2010). *World Without Oil.* http://www.worldwithoutoil.org/.

Montfort, Nick, and Ian Bogost. (2009). *Racing the Beam: The Atari Video Computer System.* The MIT Press.

Robertson, Charles J. (2006). *Temple of Invention: History of a National Landmark.* Scala Publishers.

When WordHoard Met Pliny: Breaking Down of Interaction Silos Between Applications

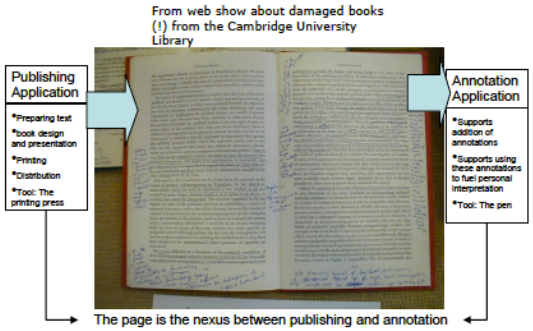
Bradley, John
john.bradley@kcl.ac.uk
Center for Computing in the Humanities, King's College London, United Kingdom

Hill, Timothy
timothy.hill@kcl.ac.uk
Center for Computing in the Humanities, King's College London, United Kingdom

One of the current issues within the Digital Humanities community is the wish to break down "silos" between different applications, usually based on the observation that it is difficult to bring two separately developed applications together even on kinds of data that they might ideally actually share. Scholarly annotation and notetaking has not been often been thought of as a kind of "anti-siloing" activity, however, they do involve the juxtaposition of materials and ideas from a broad range of different sources. In this context normal web pages and digital applications act a bit like silos – working against the ability, similar to what conventional notetaking provides, to juxtapose materials from different places, and making it difficult for a computer user to preserve those juxtapositions that are interesting.

In an attempt to recognise the central role of annotation and notetaking in scholarship, there has been recent activity in the Digital Humanities (DH) community to incorporate Web 2.0-like annotation services within a number of web resources. We think this is, in fact, the wrong way to go, and it is our contention that annotation, in fact, requires a significantly different approach.

We can see the problem with this "adding an annotation service" approach if we consider annotation on paper (see figure 1). When the book reader writes on the page he combines on that piece of paper two rather different applications that then must co-exist: the print media represented by the printed word and his/her annotation shown by the handwritten note. The owner, the technology and purpose of these two co-existing texts – the annotation and the print material – are quite different. Furthermore, whereas the printed text represents an endpoint in the "publishing application" that put it there, the hand-written annotation represents the beginning of an act of interpretation that is likely to continue into the future. In some senses, then, a printed page with an annotation on it represents a nexus between these two quite different applications: the presentation of the print, and the support for the annotation made by the individual reader. The oddness, then, of the annotation-as-a-resource service in a web resource is brought into clearer focus when we realise that if handwritten annotation on a printed page worked in the way that an annotation service on a website would operate , it would need to be a service of the book's publisher – something that would seem very peculiar, and, indeed perhaps strikingly inappropriate.



a printed page as the nexus between applications

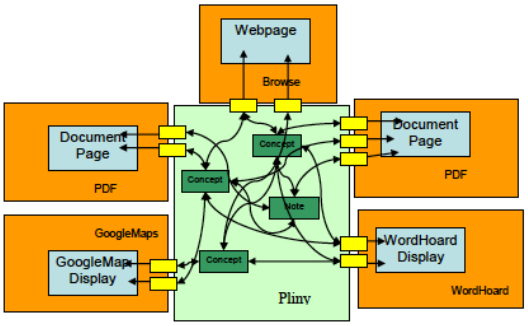
Pliny (Pliny 2009), as initially installed supports annotation for web pages, images and PDF documents. In each of these cases separate mini-applications (one for each data type) supports, simultaneously, mechanisms to display the object (web page, image or PDF) and to support annotation of these objects. Annotation items, although appearing on the web or PDF page display are also objects that work in the larger Pliny context as objects in their own right. Thus, like the printed book, the Pliny screen becomes the nexus between the "display application" of the web or PDF page and the separate-but-linked "annotation/notetaking application". Furthermore, the Eclipse platform (Eclipse 2010) on which Pliny operates already supports the dynamic addition of new applications into an existing installation. Pliny could thus be relatively straightforwardly extended to add support for annotation on other media such as video or audio, and here again the integration between these media and Pliny notes would be similar to that provided in base Pliny.

The purpose of this poster, however, moves on to a next step in the process beyond the support of other digital media. Pliny is already equipped to support annotation of relatively fixed objects such as

PDF documents or digital movies. The digital world, however, is not in fact restricted to the presentation of relatively fixed media objects but extends to supporting user interaction with dynamic applications which are less comfortably considered documents and more likely to be thought of as tools. Thus, even in the confines of the WWW with its strong document focus, DH work often explores how to stretch this document-focus in the browser to deliver a tool to the user instead. Indeed, work as diverse as TaPOR (2008) and PASE Domesday (2010) illustrate this straining at the constraints that are imposed by the document orientation.

As a part of Pliny's so-called "second agenda" (see Bradley 2007 and 2008), we have also been interested in exploring the boundary between document and application/tool in the context of annotation, and we are doing this work from within the Eclipse plugin framework where application/tool building is, in fact, an entirely natural thing. Furthermore, unlike other application building frameworks, such as Java Swing, the plugin framework supports the kind of intimate interaction and co-existence that annotation requires. In this way, Pliny's use of Eclipse's plugin framework allows us to think of how to support note-taking not only against fixed digital media, but also against the dynamic results generated by digital tools.

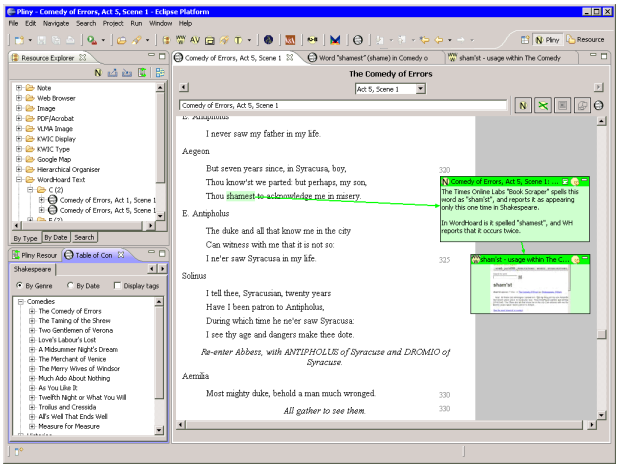
Figure 2 shows the idea in a stylised form. Here we see the Pliny application (shown here in green) co-existing with other applications. Two of them (Pliny's integrated Web browser, and Pliny's PDF annotator) are shown here as orange boxes presenting their particular digital media objects: a page in a web browser (at the top), and two PDF documents. The annotations to these objects are represented here as the little yellow boxes which, simultaneously, would actually be visually integrated with their web or PDF pages, but are also owned by Pliny. The Pliny user can also use these annotations in the Pliny application as sources for ideas that inspire him/her to construct new concepts in his/her interpretation of these materials. Pliny can represent these new concepts too, and they are represented in the green Pliny application box as dark green boxes. Thus, a Pliny item can appear both as an annotation on its media and also simultaneously as a note participating in concepts that belong to an interpretation that Pliny's user is developing.



Annotation as a kind of "glue" between applications

The "2nd agenda" part of Pliny's environment is shown by the bottom two boxes placed on both sides of the green Pliny box. Here, the objects being annotated don't come from Pliny's initial built-in applications supporting Web pages and PDF files, but from two applications that have been added: a Google maps annotation tool, and an implementation of WordHoard (WordHoard 2010) which represents a more dynamic application than the other ones do. The user has attached annotations to displays created by both these applications, and these annotations, which represent observations the user has made while s/he uses these tools, are Pliny objects (like those attached to the PDF pages) and can therefore also participate in the work done by the user within Pliny. Just like notes attached to PDF pages or web pages, notes attached to displays by Google maps or Wordhoard can also contribute materials to the user's growing interpretation.

We chose to use the funds provided by the Andrew W. Mellon Foundation's MATC 2008 award to Pliny (which hereby we acknowledge with thanks!) as a way to fund a serious exploring of this idea. With the cooperation of Martin Mueller and others at Northwestern University we have been exploring what their WordHoard tool (WordHoard 2010) would be like if it was presented in an intimately linkable environment where Pliny operates rather than as a conventional Java application in which it was originally conceived. You can see a result of this kind of interaction through our WordHoard prototype in figure 3. Not only is a note about the word "shamest" displayed attached to its occurrence in Shakespeare's Comedy of Errors, but there is also a reference to a WWW site that has nothing to do with WordHoard, but that also contains the play's text (with the word spelled differently).



WordHoard and Pliny operating as integrated applications

In conjunction with the work on WordHoard, we are also extending Pliny's code to exploit Eclipse's ability to allow the user to install a new plugin dynamically from remote online repositories. When dynamic installation has been added in Pliny. a user will be able to start with only basic Pliny, but at some point when they need it be able to add into their environment, say, our WordHoard application. We expect it will in the end be no more complex to do this in Pliny than it is to, say, add Zotero to Firefox. At our poster session the visitor will be able to see the Wordhoard integration with Pliny and with other independently built tools that can interoperate in the Pliny environment. We will illustrate what the experience is for the user to experience this kind of integrated environment. For the development community we hope to speak about the coding work that this represents and what lessons can be learned from our experience.

References

Bradley, John (2007). '"Pliny: Making a contribution; Modularity, Integration and Collaboration between Tools in Pliny." Peer reviewed poster.'. *Digital Humanities Conference 2007.* Urbana-Champaign: University of Illinois., 2007. `http://pliny.cch.kcl.ac.uk/docs/Illinois-Poster.pdf`.

Bradley, John (2008). '"Playing together: modular tools and Pliny." Draft of paper.'. *Digital Humanities 2008.* (University of Oulu, Finland, June 2008. `http://pliny.cch.kcl.ac.uk/docs/oulu-paper.html`.

Eclipse homepage. `http://www.eclipse.org/` (accessed 2010).

PASE Domesday. `http://domesday.pase.ac.uk/` (accessed 2010).

Pliny: A Note Manager. `http://pliny.cch.kcl.ac.uk/` (accessed 2009).

Text Analysis Portal for Research: TAPoR. `http://portal.tapor.ca/portal/portal` (accessed 2008).

WordHoard: An application for the close reading and scholarly analysis of deeply tagged texts. `http://wordhoard.northwestern.edu/userman/index.html` (accessed 2010).

The Wellcome Arabic Manuscripts Project

Brey, Gerhard
gerhard.brey@kcl.ac.uk
Centre for Computing in the Humanities, King's College London, UK

This poster and software demonstration will present the Wellcome Arabic Manuscripts project. The aim of this project was to create a freely available online catalogue of ca. 500 mainly medical manuscripts written in Arabic that are preserved in the Wellcome Library (London, UK). Apart from the actual manuscript catalogue the outcomes of this project that are of particular relevance to the Digital Humanities are:

- a TEI/ENRICH based XML schema adapted to meet the requirements of cataloguing Arabic manuscripts and in particular to accommodate detailed codicological and textual descriptions
- an open source, web-based, customizable cataloguing tool
- an online research tool that gives users free access to a wealth of metadata and digitized page images of the manuscripts in this collection
- high-quality digital images of each manuscript page linked to rich descriptive metadata.

The project was partly funded by the Wellcome Trust and partly by a grant from the UK's JISC Islamic Studies Programme [Henshaw, 2009]. It is a collaboration between three institutions:

1. Wellcome Library, London, UK[1]
2. Bibliotheca Alexandrina, Alexandria, Egypt[2]
3. King's College London, London, UK[3]

The Wellcome Library provided the cataloguing expertise and methodology, the digital images, and overall project management, while the web based tools were developed and are hosted by the Bibliotheca Alexandria. King's College London (Centre for Computing in the Humanities) managed the technical requirements specifications for the cataloguing tool, and adapted the TEI/ENRICH XML model[4].

The cataloguing methodology and the approach to this project was guided by Wellcome's tripartite approach to cataloguing Oriental manuscripts. This approach suggests that a manuscript should be considered as a product of craftsmen, authors and readers and therefore its production, intellectual content and subsequent use. The description of a manuscript is carried out having these three aspects in mind, viewing the manuscript:

1. as a museum object (palaeography, codicology)
2. as an intellectual creation (texts)
3. under historical user aspect (provenance, owners, editors, etc.)

As a reflection of this approach the manuscripts were catalogued in much more detail than a typical library manuscript catalogue, particularly in the areas of the materiality of the manuscript and the description of the textual content. In practical terms this approach -- together with the newly created tool described below -- allowed the cataloguing tasks to be allocated according to competence and practicality. The detailed codicological description, best done on the actual physical object, was carried out by conservators at the Wellcome Library, whereas the detailed description of the textual content (down to the transcription of chapter headings) was undertaken by cataloguers at the Bibliotheca Alexandrina.

A key objective of the project was that the encoding model should follow established standards to ensure interoperability of the manuscript catalogue and to make it extendable and as flexible as possible. A model based on TEI[5] and the format used by ENRICH[6] (a European manuscript cataloguing project) was chosen [Pierazzo, 2010]. The basic TEI/ENRICH model had to be adapted and customized to accommodate the needs of the tripartite approach. It had to be ensured, for example, that the model is compatible with standards such as Ligatus[7], an emerging standard for the detailed description of book binding features. Other extensions to the model were needed to enable the description of the very detailed physical features from a conservator's point of view, such as flaps, endbands, or covers. From a more palaeographical perspective various features relating to the scribe had to be added, for example an element to describe the Mistara, an impression on the paper achieved by applying a kind of stamp that indicated the lines a scribe would then write on. Other important palaeographical features that had to be represented were the coefficients calculated by using the "Pace" method, a system devised by Nikolaj Serikoff [Serikoff, 2001] to measure certain features of an Arabic script, such as angles of letters, or the ratio of connected and unconnected letters on a line. These features taken together are quite unique for a scribe or scribal

school and help to situate a manuscript chronologically and geographically. In order to adequately represent the intellectual content of the manuscript, further adaptations were made. In order to fully represent compound Muslim names, for example, fields for their constituent parts (e. g. patronymic, honorific) were introduced. Additional incipit-like elements were also needed to hold those formulaic passages at the beginning of Arabic manuscript texts (the Basmallah, or Tahmid that superficially look like invocations, but that also tell about the subject area of the text to follow, or the origin of the author.

In addition to the cataloguing effort by the cataloguers at the Bibliotheca Alexandrina, all of the technical development was undertaken by the Egyptian partners. The two outcomes of this substantial development effort are the web based cataloguing tool for data input and administration by Wellcome and Bibliotheca Alexandrina cataloguers and the online research tool that enables scholars and the wider public access to the rich material via browsing and searching.

The web-based tool was designed to allow the cataloguing of manuscripts into valid TEI XML files without prior specialist XML knowledge. This was achieved via the development of a Schema driven editor (SDXE) together with a "configuration grammar", the XML Skeleton Annotations (XSA) [Aboulnaga, 2010].

The XSA system automatically builds JSF (JavaServer Faces) based XML editors. These editors in turn produce schema compliant XML files that follow a certain XML skeleton. To enable users to author such XML files, the system generates web forms with fields for each data holding XML element in the skeleton. It does this by reading an XML Skeleton Annotations file (`XSA.xml`) that contains definitions for each location in the XML skeleton including a label, a help text, authority lists, user access rights, and various other information used by the system to generate the web forms. The XSA system uses the Schema Driven XML Editor to generate schema compliant XML files. The steps necessary to generate a website using XSA are as follows:

- authoring of an XSA.xml configuration file
- optional creation of facelet templates for look and feel
- creation of a blank XML record template

The central component of the cataloguing tool is the XSA configuration file. By changing this file it is possible to adapt the cataloguing tool for any other similar manuscript cataloguing tasks.

The web based research tool, i.e. the online manuscript catalogue is directed towards both specialist and non-specialist use. This means that it has to provide functionalities that address scholarly users, palaeographers, conservators, but also a wider audience whose specialist fields lie in other areas. The research tool therefore offers entry points into the repository via multiple levels and access routes, such as browsing (alphabetic or faceted) or searching (simple or advanced) for a wide ranging set of criteria. The advanced browsing and searching mechanisms take into account features peculiar to manuscripts in Arabic, for example search by stem or search by root, differentiated again by normalized character forms or defective (i. e. dotless) character forms.

During the poster session we will give a live demonstration of the web based cataloguing tool and the online research tool. This will showcase the key features of both applications and highlight the problems that were encountered and how they were overcome

2. Contributors to the Project

2.1. Wellcome Library

- Dr. Richard Aspin
- Dr. Christy Henshaw
- Dr. Nikolaj Serikoff

2.2. Bibliotheca Alexandrina

- Prof. Magdy Nagi
- Dr. Noha Adly
- Younos Aboulnaga

2.3. King's College London

- Simon Tanner
- Dr. Elena Pierazzo
- Gerhard Brey

References

Aboulnaga, Y. (2010). 'XSA - XSD to configurable JSF UI'. *software package. documentation wiki.* `http://sourceforge.net/projects/xsa-xsd2jsf/http://so`

urceforge.net/apps/mediawiki/xsa-xsd2jsf/ (accessed October 2010).

Henshaw, C. (2009). 'JISC funding for the Wellcome Arabic Manuscript Cataloguing Project'. *Wellcome Library Blog.* . http://wellcomelibrary.blogspot.com/2009/08/jisc-funding-for-wellcome-arabic.html (accessed October 2010).

Pierazzo, E. (2010). 'On the Arabic ENRICH schema'. *Wellcome Library Blog.* . http://wellcomelibrary.blogspot.com/2010/08/guest-post-elena-pierazzo-on-arabic.html (accessed October 2010).

Serikoff, N. (2001). 'Image and Letter: "Pace" in Arabic Script (a Thumbnail Index as a Tool for a Catalogue of Arabic Manuscripts. Principles and Criteria for its Construction)'. *Manuscripta Orientalia.* 7: 56-66.

Notes

1. http://library.wellcome.ac.uk/
2. http://bibalex.org/
3. http://www.kcl.ac.uk/cch/
4. http://library.wellcome.ac.uk/arabicproject.html
5. http://www.tei-c.org/
6. http://enrich.manuscriptorium.com/
7. http://www.ligatus.org.uk/

The Canadian Writing Research Collaboratory: Infrastructure Development through Partnership

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

This poster will outline the strategies for the collaborative infrastructure development, still in progress, of the Canadian Writing Research Collaboratory (www.cwrc.ca), and the results of these strategies to date. This project is funded by the Canadian Foundation for Innovation to provide an open web-based environment to foster the use of digital tools and resources for literary studies in and about Canada. At this point when Canada's literary heritage is moving online, management of information about Canadian literary and cultural history still relies on tools derived from print models, which cannot accommodate the explosion of online materials. CWRC is predicated on the assumption that literary studies will increasingly shift from the model of solitary scholars working on small groups of texts, towards working in increasingly collaborative research environments, whether as individuals or in teams, on larger sets of texts. Its infrastructure is designed to help literary scholars make better use of digital tools in their use by providing an online environment developed in response to current research in the scholarly community.

CWRC is designed around partnerships, whether one is referring to: system components; the aggregation and federation of primary and secondary materials; or the scholarly activity for which the infrastructure is designed. This project's infrastructure is a combination of computing hardware, software, and personnel to deploy a unique platform, a collaboratory, comprised of two major elements, a database and a toolkit, linked through a web-based service-oriented architecture. This poster outlines the key partnerships built into CWRC's development plan, using visual diagrams to communicate both the system architecture and the various types of partners and the relationships CWRC has with the various partners involved in the project.

The infrastructure will launch with its repository database, Online Research Canada, or ORCA, already

populated with a body of open access “seed” data contributed by partners from existing data currently residing in silos. Some of these projects are long-finished and some ongoing. Federation of data in related partner projects will expand the body of initial materials, giving the Collaboratory both a set of materials on which to test its functionality and a critical mass of data necessary to demonstrate the potential of the new research environment to scholars. A number of partnerships with contributing and pilot projects will ensure that further data is ingested while the technical infrastructure is in its early development stages.

The CWRC environment will be designed to empower online modes of research and collaboration for a community that is not well versed in such collaboration. Anticipated components of the CWRC toolkit include: tools to edit and annotate materials in and (in the case of annotating) beyond ORCA; aids to discovery, mining, and visualization of data; and collaboration and social networking tools. However, with the exception of search engines, none of these tools have been adopted broadly within the community of literary scholars, so partnering with scholars to assess various tools and select amongst

the various possibilities is key. A new scholars group will be crucial to ensuring that we anticipate as far as possible the needs of digital adepts, as well as those of established scholars. Partnerships with established tool providers such as the TAPoR portal and with research projects prototyping second-generation tools for literary studies are key to ensuring the right balance of well-tested and emergent tools. More diffuse partnerships such as use of open-source software are also fundamental to the design of this project. CWRC is extremely fortunate to be partnering with researchers in computing science and other technical fields in the development of some crucial areas such as role and workflow management, social networking, and search, federation, and data mining.

Interface will be a major challenge, since its design will likely have a greater impact than anything else on the degree of scholarly adoption. Our strategy here is to develop first the fundamental components of the system—the repository, the editor, and a role and workflow management system—with very minimal interfaces, and then to work with scholars in our contributing and pilot projects to develop an interface in response to user needs, through an iterative process.

One final, essential element in our infrastructure development is working towards long-term sustainability for the data that scholars will be entrusting to CWRC. Partnership with our host institution’s research library for long-term data curation and preservation is fundamental to that planning, and has had a shaping impact on our infrastructure development.

The poster will outline the results so far of this development strategy, since we will by the time of DH2011 be in the thick of development work. In so doing it will assess the challenges and advantages associated with CWRC’s partner-oriented strategy in relation to each of these areas, and assess the impact of this approach on both infrastructure planning and progress to date.

Sample of literature that will be used to contextualize this poster

---

References

Brindley. Lynne (2002). 'The Future of Libraries and Humanities Research: New Strategic Directions for the British Library'. *Libraries and Culture.* 37.1: 26-36..

Borgman, Christine L. (Forthcoming). 'The Digital Future is Now: A Call to Action for the Humanities'. *Digital Humanities Quarterly.* .

Cunningham, Leigh (2010). 'The Librarian as Digital Humanist: The Collaborative Role of the Research Library in Digital Humanities Projects'. *University of Toronto Faculty of Information Quarterly.* 2.2.

Deegan, Marilyn (2006). *Text Editing in a Digital Environment Follow-up Seminar Report. AHRC ICT Methods Network*

Hughes, Lorna. *Digital Tools Development for the Arts and Humanities Report. AHRC ICT Methods Network*

Martin, Shawn (January 2007). *Digital Scholarship and Cyberinfrastructure in the Humanities: Lessons from the Text Creation Partnership. Journal of Electronic Publishing.* .

Siemens, Lynne (2009). *‘It's a team if you use “reply all” ’: An exploration of research teams in digital humanities environments. Literary and Linguistic Computing.* .

Sustainability of Digital Resources in the Arts and Humanities Expert Seminar Papers. AHRC ICT Methods Network

Ragaz, Sharon. *Text Editing in a Digital Environment Rapporteur's Report. AHRC ICT Methods Network*

Unsworth, John (2006). *Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences.* New York: ALCS.

Zorich, Diane M. (2007, 2008). *A Survey of Digital Humanities Centers in the United States. Prepared for the Council on Library and Information Resources (CLIR).* .

Discovering Citation Relations among the Imperial Court Documents of Qing China

Chen, Shih-Pei
gail@turing.csie.ntu.edu.tw
Department of Computer Science and Information Engineering, National Taiwan University

Ho, Hou-Ieong
brent@turing.csie.ntu.edu.tw
Department of Computer Science and Information Engineering, National Taiwan University

Tu, Hsieh-Chang
champiyen@turing.csie.ntu.edu.tw
Department of Computer Science and Information Engineering, National Taiwan University

Hsiang, Jieh
jhsiang@ntu.edu.tw
Department of Computer Science and Information Engineering, National Taiwan University; Research Center for Digital Humanities, National Taiwan University

---

The dynasties of imperial China had always had sophisticated governing systems to run the vast empire (Chien, 1952). While many of these dynasties left large quantities of imperial court documents, the last dynasty, Qing (1644-1911), produced the largest volume. These documents have been a major source of primary research material for studying Qing era China since they provided the most direct and first-hand details of how national affairs were handled. Among them, two of the most important kinds are *Imperial Edicts* (from the emperors to his officials) and *Memorials* (reports from officials to the emperor). The number of Memorials increased significantly after Emperor Kang-xi (康熙 – reigned from 1662 to 1723) allowed senior *local* officials to report to him directly (Chuang, 1979). The ability for the emperors to obtain first-hand information directly from local officials was among the major reasons why the Qing imperial courts did not suffer as much interference from people surrounding the emperors, such as eunuchs and family members of the empress dowagers, as in the previous Ming Dynasty.

Qing Dynasty had a systematic way to archive official documents. However, although most of the archives were organized chronologically, the court documents

involved in a specific event might span several months and were often kept in different archives. For instance, if the emperor received a Memorial reporting a rebellion in some province, he might decide to issue an Imperial Edict to give instructions to relevant officials. The Memorial, depending on its character, might be kept (or had copies made) in the Archives of the Imperial Palace (宮中檔), Archives of the Grand Council (軍機處檔案), or the Grand Secretariat Archives (內閣大庫). The Edict might have records in the Imperial Decrees Archives (上諭檔), Archives of the Diary-Keepers (月摺檔, 起居注), or the Archives of the Imperial Palace (宮中檔), or the Grand Secretariat Archives mentioned above. Worse yet, these archives are now kept at different locations, notably the National Palace Museum (National Palace Museum, 2001), the Institute of History and Philology of Academia Sinica (Institute of History and Philology of Academia Sinica, 2001), both in Taipei, and the First Historical Archives of China (FHAC, 1995), in Beijing. Although digitization effort at the former two institutions made these archives easier to access than before, it remains a cumbersome task to collect documents covering the same event and rebuild their original linkage.

In this paper, we present an approach to restore an important relation: the citation links among the imperial court documents. To be more precise, a Memorial from an official often quotes earlier Imperial Edicts as the directive for the activities that he is reporting. On the other hand, an Imperial Edict may also cite earlier Memorials as the reason for issuing the decree. In an important historical event such as the Taiping Rebellion, there may be hundreds of Imperial Edicts and Memorials that form a complex web of successive citations. We call such a graph an IE-M diagram. Figure 1 is an illustration.

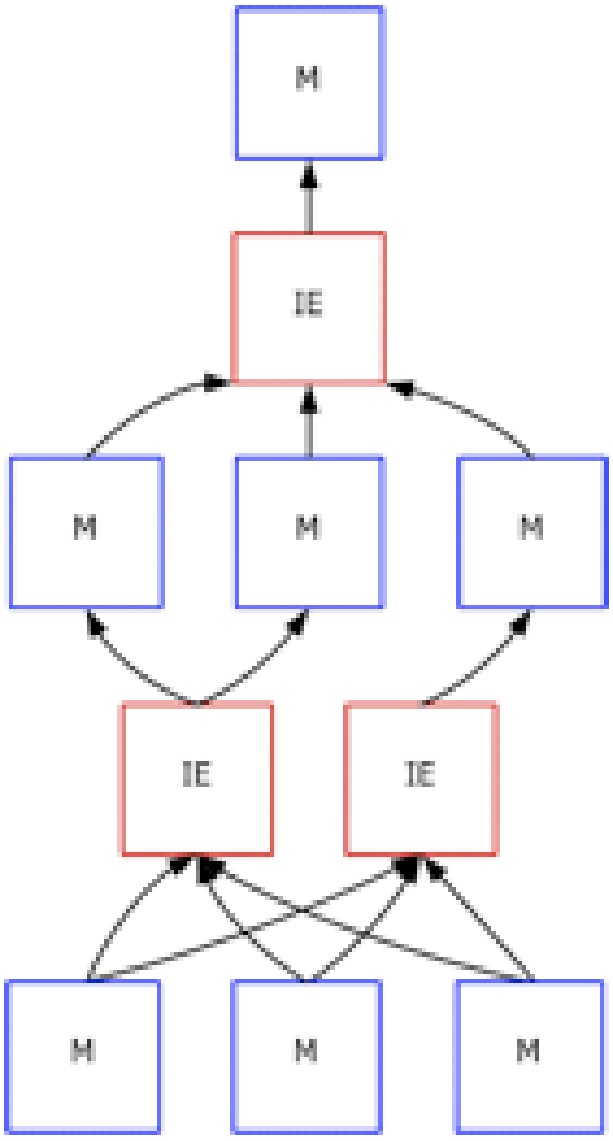


Fig. 1 An IE-M Diagram

To discover a citation relation we need to first detect whether a document has cited previous documents. This problem is similar to plagiarism detection (or copy detection), which is to detect whether a part of a document is copied from other materials without acknowledging the source (Shivakumar and Garcia-Molina, 1995; Si, Leong and Lau, 1997; Timothy and Justin, 2003). However, unlike plagiarism detection for which an exhaustive comparison among documents might be necessary, in our case there are often specific phrases occurred around quotations, which we call *syntactic anchors*. In the case of a Memorial citing an Imperial Edict, the former usually contains an anchor that starts with *adhering to the Imperial Edict* (奉上諭) and ends with *By the Emperor Himself. That is all* (欽此). The text in between are quoted verbatim from the Edict (although usually not in its entirety) (see Fig. 2).

Fig. 2 In a memorial, the sentences in between the syntactic anchors are text quoted verbatim from the source edict.

The anchors involved in an Imperial Edict citing a Memorial may have a number of varieties. The quotations may also be done in a more casual manner, since the emperors did not feel obliged to quote carefully. After identifying the anchor and the quoted text, our method extracts a segment of the latter (called *signature*) and applies a text-matching algorithm to see if it appears in any document in the database. A document of the right type that contains the signature is called a *candidate*.

We remark that because the quoted text may not be very precise, a certain degree of fuzziness is incorporated in our algorithm so that minor differences can be tolerated. A detailed algorithm is listed in Fig. 3.

For signature s of a Memorial d and for d′ in the database (d′ is other than d), express s as a string (sequence) of N words:

$$s = \langle w_1, w_2, \ldots, w_n \rangle$$

We say that d′ is a candidate if

(1) d′ contains a substring: $\langle w_1, w_2, \ldots, w_n \rangle$

(2) d′ contains a substring s′, where s and s′ share >= N-R common subsequence. (R is a predefined integer indicating the number of different words to be tolerated.)

In our experiment, we set N=15 and R=1.

Fig. 3 The text-matching algorithm used for Memorials citing Imperial Edicts.

The comparison produces a list of *document, signature, candidate* tuples. We then use metadata to filter unlikely tuples, and present the findings to historians for manually validation to ensure accuracy. The overall process of our approach is shown in Fig. 4.

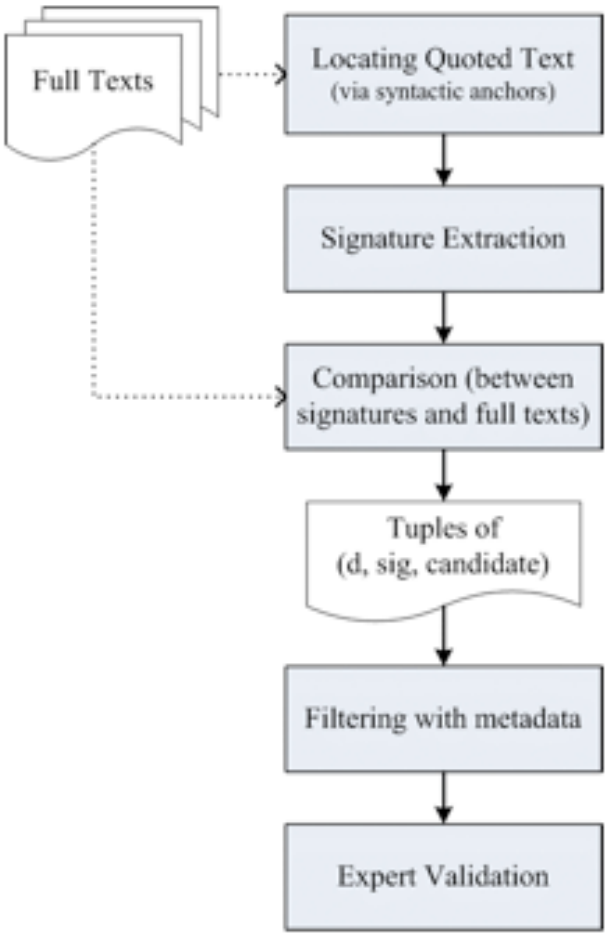


Fig. 4 The overall process of our approach.

We applied this method to the “Collection of Taiwan-related Court Documents from the Ming and Qing Dynasties” corpus in THDL, the Taiwan History Digital Library (Chen, Hsiang, Tu, and Wu, 2007; Research Center for Digital Humanities, 2009). This corpus contains 37,831 imperial court documents of mostly Qing era that are related to Taiwan. The documents, mainly Imperial Edicts and Memorials, were selected from 235 different sources, including the archives mentioned earlier in this paper. They were chosen by historians, typed as full text, punctuated, proof-read, and were supplemented with metadata records (Wu, 2004; Chiu, 2006).

Using the method described above, we discovered 5,403 pairs of citation relations from these documents, among which 3,947 pairs are Memorials citing Edicts, and 1,456 pairs are Edicts citing Memorials. By taking the transitivity on the discovered citation pairs, we produced 1,258 IE-M diagrams (see Table 1), the largest of which involves 152 documents.

Graph size	Main Content
152	Edicts and Memorials dealing with the Lin Shuangwen Rebellion, the most severe revolt in Taiwan (1786-87) during Qing reign
88	Edicts and Memorials concerning the incident of Nerbudda and Ann, two British vessels sank off the coast of Taiwan during the First Opium War (1839-42)
85	Edicts and Memorials about the Mudan She (□□□) incident, an expedition launched by the Japanese in retaliation for the killing of 54 Ryukyuan sailors by Paiwan aborigines near the southwestern tip of Taiwan. (1874-75)
65	Edicts and Memorials regarding the Keelung and Tamsui invasions and the subsequent blockade in northern Taiwan by the French during the Sino-French War (1884-85)
57	Edicts and Memorials about the raids of Cai Qian (□□), a Chinese pirate, and his subsequent capture (1806-09)

Table 2 The main content of the IE-M diagrams sized over 50. All of them are about major events in the Taiwanese history.

Fig. 3 is an example of such a diagram, in which the blocks at the top and bottom are Imperial Edicts while the middle three are Memorials. The arrows between the blocks indicate citations.

After examining the diagrams, we found that all of the larger ones (size over 50) are about major events in Taiwanese history (see Table 2).

Graph size	Number of graphs
152	1
80-89	2
60-69	2
50-59	3
40-49	1
30-39	5
20-29	10
10-19	69
5-9	261
2-4	902
Total	**1,258**

Table 1 A summary on the sizes of the 1,258 IE-M diagrams produced by our method.

Tracing through the citations shows the process of how the Qing Imperial Court dealt with crises occurred in Taiwan, a far-flung island of the vast empire. For example, the largest IE-M diagram illustrated how Emperor Qianlong (乾隆) handled the Lin Shuangwen (林爽文) Rebellion, the largest civil unrest in Taiwan during the Qing reign. The diagram vividly reflected how the rebellion, first dismissing as a minor local disturbance, developed into an island-wide revolt. (The rebels even overran a prefectural seat and had another under siege for more than six months). It also showed how the local officials, failed at suppressing the revolt, pointed fingers at each other or reported false victories. Qianlong finally realized the severity of the situation and sent Fukangan (福康安), one of his most trusted generals, to put down the rebellion. (Qianlong himself considered the pacification of Lin Shuangwen Rebellion one of his “Top Ten Military Achievements”.)

In this paper we described an approach to discover citation relations among Imperial Edicts and Memorials of the Qing Dynasty. The transitivity closures of the relations are captured in a concept of *IE-M diagram*, which reveals how a historical event developed through the correspondences between the Qing imperial court and the local governments. Our method demonstrates how to use information

technology to explore and identify important relations among historical documents that would be hard to find otherwise. We applied this method to THDL, Taiwan Historical Digital Library, and found 1,258 such diagrams. We should remark, however, our method can be applied to other and larger corpuses provided that the full text of the documents are available. We are currently working with historians to explore other significant relations to which our method can be applied.

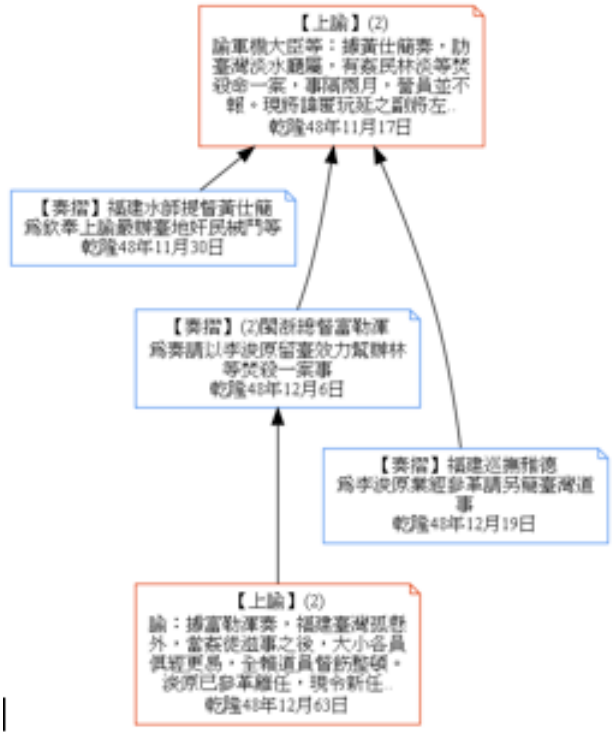


Fig. 5 An example of IE-M diagram.

References

Chen, S.P., Hsiang, J., Tu, H.C. & Wu, M.C. (2007). 'On Building a Full-Text Digital Library of Historical Documents'. *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, Lecture Notes in Computer Science no. 4822.* Goh, D.H.L., ed.. New York: Springer Berlin,, pp. 49-60.

Chien, M. (1952). *Chung-kuo li tai cheng chih te shih.* Hong Kong.

Chiu, W.J. (2006). 'The Digital Project of Taiwan-Related Archives in Ming and Qing Dynasty'. *The Library Yearbook of ROC 2006.* Taipei: National Central Library.

Chuang, J.F. (1979). *Qing dai shi liao lun shu.* Taipei: National Place Museum.

The First Historical Archives of China (FHAC) (1995). *Guangxu chao zhu pi zou zh.* Beijing: Zhonghua shu ju.

National Palace Museum. *The Archives of the Grand Secretariat in Academia Sinica* . `http://archive.ihp.sinica.edu.tw/mctkm2/index.html` (accessed 10 March 2011).

Research Center for Digital Humanities of National Taiwan University. *Taiwan History Digital Library.* `http://thdl.ntu.edu.tw` (accessed 10 March 2011).

Shivakumar, N. and Garcia-Molina, H. (1995). 'SCAM: A Copy Detection Mechanism for Digital Documents'. *The 2nd International Conference in Theory and Practice of Digital Libraries..* Austin, Texas, 1995.

Si, A., Leong, H.V. and Lau, R.W.H. (1997). 'CHECK: a document plagiarism detection system'. *Proceedings of the 1997 ACM symposium on Applied computing.* San Jose, California, 1997.

Timothy, C. H. and Justin, Z. (2003). 'Methods for identifying versioned and plagiarized documents. '. *Journal of the American Society for Information Science and Technology.* 54(3): 203-215..

Wu, M.C. (2004). *Taiwan shi liao ji cheng ti yao.* Taipei: Council for Cultural Affairs.

A Labanotation Editing Tool for Description and Reproduction of Stylized Traditional Dance Body Motion

Choensawat, Worawat
gr0011es@ed.ritsumei.ac.jp
School of Science and Engineering, Ritsumeikan University

Takahashi, Sachie
nr013082@ed.ritsumei.ac.jp,
School of Letters, Ritsumeikan University

Nakamura, Minako
nakamuraminako@gmail.com
Graduate School of Humanities and Sciences, Ochanomizu University

Hachimura, Kozaburo
hachimura@media.ritsumei.ac.jp
School of Science and Engineering, Ritsumeikan University

1. Introduction

A stylized traditional dance has uniqueness in itself which reflects history, culture, emotion expression, etc. When recording and representing this traditional dance body motions, it is important to have capabilities for handling these very characteristic body movements, which can probably be handled with the full-set of Labanotation (Hutchinson, 1997). However, notation score becomes extremely complicated, and we will not be able to comprehend what is that movement.

We are facing a problem how to realize a method of describing detailed features and nuance of artistic, traditional dance movements while suppressing the complexity in notation score.

Several graphics applications has been developed for preparing Labanotation scores (Brown and Smoliar, 1976; Fox, 2000) and generating body movement (Calvert, 2007; Coyle *et al.* 2005; Coyle *et al.* 2002; Wilke *et al.* 2005). However none of them solved the abovementioned problem yet.

We have been working on a system named LabanEditor (Kojima *et al.* 2002; Nakamura and Hachimura, 2006). It includes functionalities of both inputting/editing Labanotation score and displaying character animation so that beginners who are not familiar with Labanotation can study about its description by a trial-and-error approach.

In this paper, we aim at description and reproduction of the body motion of stylized traditional dances by using fundamental elements of Labanotation while keeping the quality of body movement of CG character animation. We propose and implement a dynamic template technique enabling users to notate stylized traditional dances and reproducing it in 3D CG animation from a Labanotation score.

2. Labanotation

Labanotation is a graphical notation system for recording human body movements. Figure 1 (a) is an example of Labanotation scores corresponding to dance motion. A Labanotation score is drawn in a form of vertical staff where each column corresponds to a body part. Figure 1 (b) shows the basic arrangement of columns in the staff. The horizontal dimension of the staff represents the parts of the body, and the vertical dimension represents time.

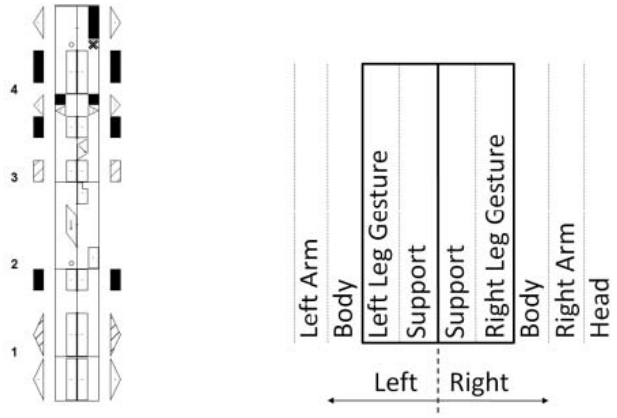


Figure 1. Labanotation scores; (a) Example of Labanotation scores and (b) Columns of Labanotation representing body parts.

3. LabanEditor

LabanEditor (Kojima *et al.* 2002) is an interactive graphical editor for editing Labanotation scores and displaying the 3D CG character animation associated with scores.

We added the new features to LabanEditor as follow:

1. Dynamic template technique enabling users to notate movements and reproducing it in 3D CG animation using fundamental description of Labnotation.

2. Motion control module for manipulate the motion expression among key-frames in order to make the animation more natural.

3.1. User Interface

With LabanEditor, users are able to input and edit the score and then display the CG animation immediately, which makes possible to interactively confirm the movements. Users can zoom in/out and change the viewpoint of the 3D scene on all three axes.

While replaying the Labanotation score, users can observe the animation as well as the red horizontal line cursor, moving upward corresponding to the animation progresses, as shown in Figure 2.

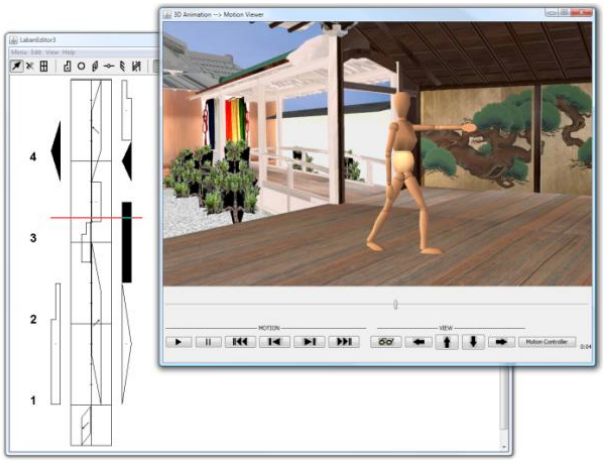


Figure 2. CG animation display window.

3.2. Generating the CG Animation

Labanotation scores can be represented as a simple format called LND (Kojima *et al.* 2002), which uses alphanumeric characters to represent basic symbols. To create 3D character animation, we have to convert LND into animation data. The format of LND representation is shown in Figure 3. The lines that begin with "#" indicate the fundamental parameters of Labanotation. The movement of a body part is specified by the line followed by a command "direction", which corresponds to the Labanotation direction symbols.

```
#rhythm        4/4
#speed         120
#unit_per_line 7
#unit_total    56
#unit   0
direction      l_support     place   mid                     standard
direction      r_support     place   mid                     standard
direction      r_arm         right   mid                     standard
direction      l_arm         left    mid                     standard
#unit   1
direction      l_support     place   mid     0.0     2.0     standard
direction      r_support     place   mid     0.0     2.0     standard
direction      r_arm         right   high    0.0     2.0     standard
```

Figure 3. Example of LND.

LND describes a pose just like key-frame body postures for animation, so that we can produce motion of the body part by simply applying interpolation between start and end key-frame poses. A key-frame pose of a body part at a timing corresponding to an end of the symbol is defined by a Labanotation symbol.

Our system converts direction symbols into animation key-frames by using a template file for a mapping between the symbol and its corresponding pose of the body part.

The template file describes the relationship between a direction symbol at the particular column and the rotation and the translation of the corresponding joint. Figure 4 shows a notation and description in a template file, and the resulting pose.

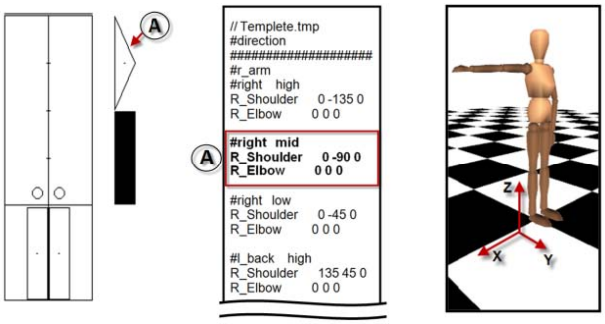


Figure 4. Relationship between user input symbols and a template file; (a) User input symbol, (b) Part of a template file, and (c) Snapshot of the CG animation corresponding to the template in (b).

3.3. Dynamic Template Technique

Due to the rough resolution of fundamental elements of Labanotation, similar but distinct poses are sometimes defined with the same symbol in a Labanotation score.

To solve this problem, we invented the method of dynamic templates in order to represent very specific movements using the fundamental subset of Labanotation symbols only. With the dynamic template technique, we can represent these characteristic motions by changing the template files dynamically

during a display process, while using very fundamental symbols.

Figure 5 (a) shows the interface of editing template file. Users can activate the editing window by double clicking on a Labanotation symbol. For example, suppose the symbol in the Labanotation score, indicated by a red color in Figure 5 (a), was selected by a user, then, the user can directly edit the joint angles on an editable template panel as shown in Figure 5(a).

Alternatively, the graphically editable template, which is activated by clicking the „Animate" button in Figure 5 (a), enables the user to edit the template by adjusting the slide bars and observe the resulting pose as shown in Figure 5 (b).

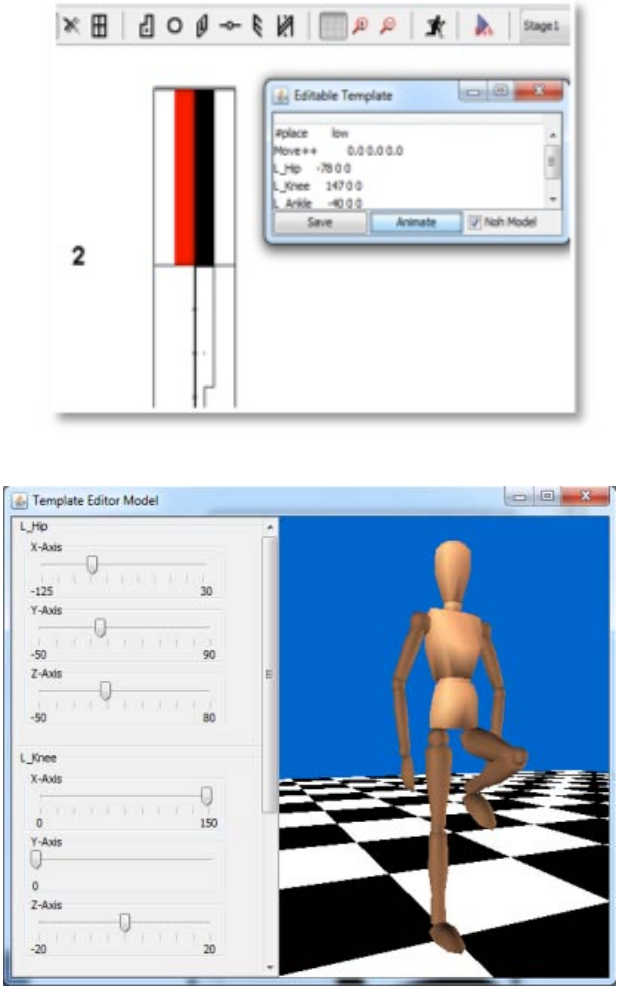


Figure 5. Interface for editing a template; (a) Template editing panel and (b) Graphical template editing panel.

The information of template files is inserted into a LND file corresponding to the start time as shown in Figure 6. The command "#include" determines the template file used at a particular timing. As a result, in this case, the Labanotation score shown in Figure 5(a) will be interpreted as the LND file shown in Figure 6.

```
#version 1
#beat    4/4
#tempo   100
#include          A.tmp
direction l_support     place   low   0.0   0.0
direction r_arm         place   low   0.0   0.0
direction r_support     back    mid   0.0   4.0
#include          B.tmp
direction r_support     place   low   4.0   8.0
rotation  l_support     place   low   4.0   8.0
```

Figure 6. Example of LND file using dynamic templates.

During the animation process, the Labanotation symbols, in format of LND, are mapped to the key-frame pose indicated by the current template.

3.4. Motion Expression Control

The motion expression control module controls the animation of character model from one key-frame to the next key-frame. We implemented a module for controlling the motion by applying a non-linear interpolation, cubic Bezier curve, in order to create natural movement.

EQUATION (1)

Where *f(t)* is an interpolated position or joint angle at time t and a normalized time scaled from 0.0 to 1.0, respectively. $P_0(0,0)$ and $P_3(1,1)$ is the start and end points, respectively, while P_1 and P_2 are the control points which can be moved freely as shown in Figure 7.

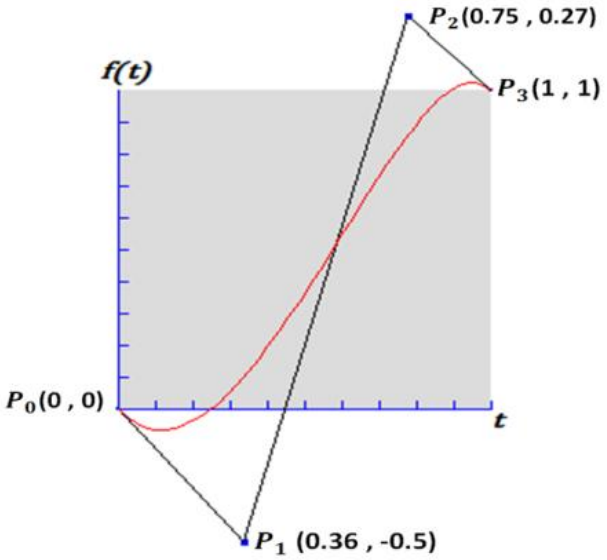


Figure 7. Motion expression controller user interface.

Figure 8 shows two snapshots of the CG animation corresponding to the motion expression graphs on the left.

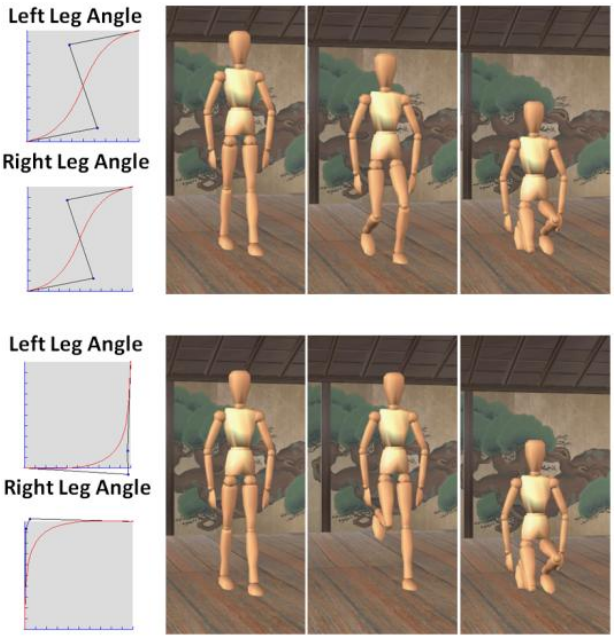


Figure 8. CG animation snapshots corresponding to the motion expression graphs.

4. Use of LabanEditor for Noh Plays

Noh is the most famous and characteristic Japanese traditional performing arts in the stylized form of a musical dance-drama that has been performed since the 14th century. There are about 240 plays in the repertoire from five Noh schools (Ortolani, 1995).

Noh body movement is Japanese highly stylized and is not the same as ordinary human movement. Therefore, the direction symbols used in Labanotation must be interpreted differently when we handle Noh plays and generate body motion from it. This has been realized by preparing motion template files which are editable to represent specific motions in that particular performance.

Snapshots of Noh motion are shown in Figure 9. Figure 9 (a) and (b) show the reproduced animation of Noh body motion from Labanotation score using the Noh templates and standard templates, respectively.

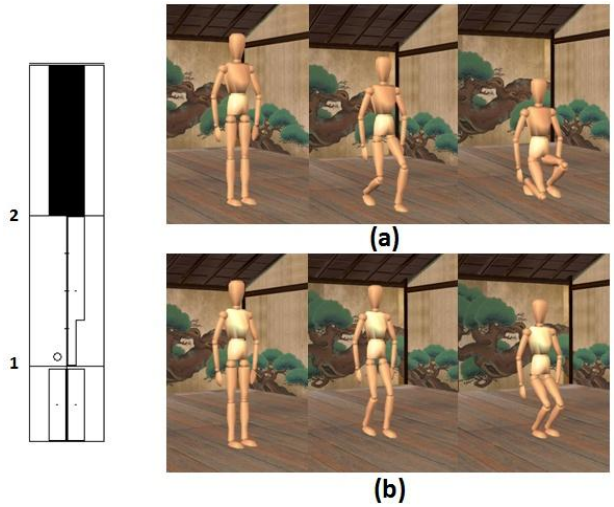


Figure 9. CG character animation of Noh Kata (Shitai)
(a) CG character animation corresponding to the Labanotation score on the left using Noh (Shitai) template
(b) CG character animation corresponding to the Labanotation score on the left using standard template.

5. Conclusions

In this paper we proposed an approach of description and reproduction of stylized traditional dances such as Noh plays with Labanotation. A new version of LabanEditor, LabanEditor3, successfully describes and reproduces Noh motions by using the dynamic template method.

We have obtained a major achievement. Our approach shows that we can describe and reproduce Noh plays with the fundamental description of Labanotation, with a limited number of symbols by using the dynamic template method.

References

Hutchinson, A (1977). 'Labanotation'. Theatre Arts Books.

Brown, M. D., Smoliar, S.W. (1976). 'A Graphics Editor for Labanotation'. *Proceedings of the 3rd Annual Conference on Computer Graphics, Interactive Techniques and Image Processing, ACM Computer Graphics.* V. 10 (2), pp. 60-65.

Fox, I (2000). 'Documentation Technology for the 21st Century'. *World Dance 2000 Academic Conference, Papers and Abstracts.* Pp. 137–142.

Calvert, T (2007). 'Animating Dance'. *Proceedings of Graphics Interface.* Pp. 1-2.

Calvert, T, Wilke , L, Ryman, R, Fox, I (2005). 'Applications of Computers to Dance'. *IEEE Computer Graphics Application.* 2: 6-12.

Coyle, M, Maranan, D, Calvert, T (2002). 'A Tool for Translating Dance Notation to Animation'. *Proceedings of Western Computer Graphics Symposium*.

Wilke, L, Calvert, T, Ryman, R, Fox, I (2005). 'From dance notation to human animation, The LabanDancer Project, Motion Capture and Retrieval'. *Computer Animation and Virtual Worlds.* 3-4: 201-211.

Kojima, K, Hachimura, K, Nakamura, M (2002). 'LabanEditor: Graphical Editor for Dance Notation'. *Proceedings of IEEE 2002 International, Workshop on Robot and Human Interactive Communication.* 59-64.

Nakamura, M, Hachimura, K (2006). 'An XML Representation of Labanotation, LabanXML, and Its Implementation on the Notation Editor LabanEditor2'. *Review of the National Center for Digitization (Online Journal).* 47-51.

Ortolani, B (1995). *The Japanese Theatre: From Shamanistic Ritual to Contemporary Pluralism.* Princeton University Press.

The Tesserae Project: Intertextual Analysis of Latin Poetry

Coffee, Neil
ncoffee@buffalo.edu
University at Buffalo, SUNY

Koenig, J.-P.
jpkoenig@buffalo.edu
University at Buffalo, SUNY

Poornim, Shakthi
poornima@buffalo.edu
University at Buffalo, SUNY

Forstall, Christopher
forstall@buffalo.edu
University at Buffalo, SUNY

Ossewaarde, Roelant
rao3@buffalo.edu
University at Buffalo, SUNY

Jacobson, Sarah
University at Buffalo, SUNY

The Tesserae Project has created a freely available web tool for analyzing text reuse (intertextuality) that automatically identifies matching two-word phrases (bigrams) in Latin poets using one of two search algorithms. Comparison with the results of traditional scholarship demonstrates the efficacy, current limitations, and potential of this approach. Automatic bigram matching by morphological form and dictionary headword detects a significant number of parallels identified by traditional methods. Results so far do not fully replicate traditional scholarship, but the incorporation of further feature sets holds the potential of approaching this standard. Bigram detection produces more systematic results, permits large-scale intertextual study, and identifies less conspicuous parallels.

2. Computational Approaches to Intertextuality

The reuse of elements from other texts has been understood as a fundamental part of textual signification from ancient Alexandria to modern times. Traditional methods of identifying specific parallels have relied upon the scrutiny and memory of scholars (Hinds 1998, Edmunds 2001). Researchers have

recently begun to employ computational means to facilitate and standardize intertextual study, as well as to open new perspectives. Two major lines of approach are phrase (n-gram) matching (e.g. Cummings 2009) and comparison of element length (Holmes 2010). In the field of classical Greek and Latin literature, the Perseus Project has identified five computationally tractable features, including bigram matches, for assessing the similarity of phrases in different texts and has offered a method for cross-linguistic phrase matching (Bamman and Crane 2008, Bamman and Crane forthcoming). A program developed by the eAQUA project locates explicit quotations in the Thesaurus Linguae Graecae corpus of Greek texts (Büchler, Geßner et al. 2010).

3. Tesserae Search

The goals of the Tesserae Project are to create a website that facilitates intertextual search of classical Latin texts (http://tesserae.caset.buffalo.edu) and to make computational methods and results accessible to traditional scholars. The Tesserae group chose bigram matching as the method most similar to the standard philological search for parallel phrases.

The tool finds similar phrases by matching two words in one text with two words in another. Users can choose two of 26 prepared texts for comparison using one of two search methods. Version 1 matches two identical words from each text, in any order, with no more than four words between them. Version 2 matches words anywhere in an individual sentence by dictionary headword using the Archimedes Project Morphology Service (http://archimedes.mpiwg-berlin.mpg.de/arch/archimedes.new.html), and employs an experimental ranking system to help the assessment of their potential significance. In both versions, the most common words are by default excluded to eliminate potentially insignificant matches. Users can modify search settings with an advanced tab.

4. Version 1 Test

A Version 1 test compared book 1 of Lucan's 1st century CE epic *Civil War* with the whole of the 1st century BCE epic *Aeneid* by Vergil. The resulting 160 parallels were ranked for significance by traditional philological analysis on a 5 (most significant) to 1 (error) scale. The ranked parallels were compared to those collated from a standard commentary on Lucan's *Civil War* (Viansino 1995). Tesserae discovered 87 results judged significant (types 5-3) as compared with 81 of Viansino. Tesserae and Viansino shared only 14 results in common. Tesserae returned results distributed more evenly through *Civil War* book 1 than Viansino, whose parallels clustered at the beginning and end of the book.

5. Version 2 Test

Version 2 eliminates errors from Version 1, matches by dictionary headword rather than exact form to account for inflection, and takes whole sentences rather than word windows as the unit of comparison. A Version 2 test again compared *Civil War* book 1 to the whole *Aeneid*, and the results were measured against the parallels given by all major Lucan commentators (Heitland and Haskins 1887, Thompson and Bruère 1968, Viansino 1995, Roche 2009). The expanded search parameters of Version 2 returned significantly more results than did Version 1: 2,994 vs. 160. Version 2 produced numbers of types 5 and 4 comparable to Version 1, but considerably more type 3s, as in the following table, where the results of commentators have been collated and graded on the same scale for comparison.

6. Conclusions

Tesserae fulfills part of its purpose by quickly generating a convenient list of possible intertextual parallels for inspection. The combined tests further demonstrate that Version 1 and Version 2 deliver comparable numbers of the types of parallels scholars have traditionally valued, close morphological similarities of non-frequent words. These results are illustrated in the following chart.

	Type 5	Type 4	Type 3	Total Significant
Version 1	19	27	41	87
Version 2	26	43	262	331
Viansino 1995	30	17	34	81
Roche 2009	85	67	170	322
All commentators*	96	81	195	372

type 5: strong verbal similarity with meaningfully analogous context
type 4: moderate verbal similarity with meaningfully analogous context
type 3: verbal similarity without substantially analogous context
*All commentators: Heitland and Haskins 1887, Thompson and Bruère 1968, and Viansino 1995, Roche 2009. This counts the total number of unique parallels found, so the same parallel found by different commentators is counted only once.

Parallels between *Civil War* 1 and *Aeneid* Found by Tesserae. Versions 1 and 2 Compared with Commentators

Although Tesserae found meaningful parallels, it did not discover the majority of those found by the commentators. Most undetected matches had features that Tesserae does not currently recognize, including similarity of location, meaning, meter, and sound. The results that Tesserae did find, however, appeared in patterns resembling those found by commentators. For example, the commentators found fewer highly significant reference types in the second half of *Civil War* 1 (type 5, 63 vs. 33 in the second half; type 4, 41 vs. 40), and the combined Tesserae results show a similar, though steeper, decline (type 5, 29 vs. 8; type 4, 30 vs. 27). Tesserae thus supports the cumulative suggestion of the commentators that Lucan establishes substantially more highly significant parallels to the *Aeneid* in the first half of *Civil War* book 1 than in the second.

Conversely, Tesserae detects more type 3 references in the first half of the poem than the second, whereas the commentators find the opposite (Tesserae: 158 vs. 122; Commentators: 89 vs. 106). One possible explanation for this difference is that commentators overlook less significant parallels when there are more significant parallels in the same vicinity. On this interpretation, the larger number of significant parallels in the first half of the *Civil War* led to a correspondingly reduced detection of less significant parallels by commentators, whereas Tesserae detected all types more consistently. Tesserae search thus serves as a complement to and check on traditional analysis in this and other respects, as when in several instances Tesserae returned more links to a given phrase than were noted by commentators.

For longer texts, the proportion of insignificant to significant parallels presented is currently too high for Version 2 to be fully useful, as substantial time is required to sort and analyze the results. Future work will involve developing a system that replicates the manual scoring used for testing to more easily identify different types of parallels. Other planned improvements include the addition of new search parameters and texts, the addition of Greek and other languages, user uploaded texts, and the ability to perform a secondary search for found phrases across a corpus.

References

Bamman, D. and G. Crane (2008). 'The Logic and Discovery of Textual Allusion.'. *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008).* Marrakesh.

Bamman, D. and G. Crane. *Discovering Multilingual Text Reuse in Literary Texts [forthcoming].*

Büchler, M., A. Geßner, et al. (2010). 'Unsupervised Detection and Visualization of Textual Reuse on Ancient Greek Texts'. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science 1*.

Cummings, James (4 September 2009). 'TEI-Comparator'. *In my <element/>..* `http://blogs.oucs.ox.ac.uk/jamesc/2009/09/04/tei-comparator/`.

Edmunds, L. (2001). *Intertextuality and the Reading of Roman Poetry.* Baltimore: Johns Hopkins University Press.

Heitland, W. E. and C. E. Haskins (1887). *M. Annaei Lucani Pharsalia.* London: G. Bell.

Hinds, S. (1998). *Allusion and Intertext: Dynamics of Appropriation in Roman Poetry..* Cambridge: Cambridge University Press.

Holmes, M. (2010). 'Using the Universal Similarity Metric to Map Correspondences between Witnesses.'. *Digital Humanities 2010 Abstracts.* `http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-693.pdf`.

Roche, P., Ed. (2009). *Lucan: De bello civili. Book 1.* Oxford: Oxford University Press.

Thompson, L. and R. T. Bruère (1968). 'Lucan's Use of Vergilian Reminiscence'. *Classical Philology.* 63: 1-21.

Trillini, R. H. and S. Quassdorf (2010). 'A 'Key to All Quotations'?: A Corpus-based Parameter Model of Intertextuality.'. *Literary and Linguistic Computing.* 269-86.

Viansino, G. (1995). *Marco Annaeo Lucano: La Guerra Civile (Farsaglia) libri I-V..* Milan: Arnoldo Mondadori.

Bamboo Technology Project: Building Cyberinfrastructure for the Arts and Humanities

Cole, Timothy
t-cole3@illinois.edu
University of Illinois, Urbana-Champaign

Fraistat, Neil
fraistat@umd.edu
Maryland Institute for Technology in the Humanities (MITH)

Greenbaum, David
dag@berkeley.edu
University of California at Berkeley

Lester, Dave
dlester@umd.edu
Maryland Institute for Technology in the Humanities (MITH)

Millon, Emma
emillon@umd.edu
Maryland Institute for Technology in the Humanities (MITH)

1. Introduction

This poster will provide an overview of the Bamboo Technology Project and demonstrate some of its prototypes. We are currently midway through an 18-month implementation phase that builds on the Bamboo Planning Project – a process that engaged approximately 600 faculty, librarians, and technologists from 115 institutions between Spring 2008 and Autumn 2010 – both of which have been funded by the Andrew W. Mellon Foundation.

The Bamboo Technology Project (BTP) is a multi-institutional, interdisciplinary effort that brings together humanities scholars, librarians, and information technologists to tackle the question: "*How can we advance arts and humanities research through the development of shared technology services?*" Comprised of an international partnership of 10 universities,[1] the project is devoted to building applications and shared infrastructure for humanities research.

2. Structured Work Areas

During the Bamboo Planning Project, we found that scholar participants repeatedly expressed their desire for technology that would allow them to annotate, to collaborate, to gather materials, to organize information, to share materials, to store and preserve materials, and to use social media. Building upon workshop discussions, we identified a more general set of categories to describe central scholarly practices in the humanities that could benefit from technology. Prominent among them were the practices of aggregation, annotation, consideration, engagement, and interaction.[2] Using these categories as guidelines, and drawing on the project's leadership and participation from different partner institutions, we organized ourselves into four major interconnected areas:

3. Structured Work Areas

- ***Research Environments*** "Adapters" are being built that will enable scholars working within the Project Bamboo Research Environments to access and apply research tools to dispersed content collections such as materials in the HathiTrust and Perseus collections. Research is being conducted to enable OpenSocial gadgets to be leveraged within the Project Bamboo Research Environments and potentially other areas of the Bamboo ecosystem. OpenSocial is a promising approach to enable us to develop functionality once and apply it many times. We are also building a Tools and Services Registry through which scholars can discover other software tools that their fellow humanist scholars are using.
- ***Corpora Space*** We will carry out a community design process to plan for a future generation of corpora-centered virtual research environments to support the central and growing importance of data and of corpora curation for humanities' scholarship. The results of this design process will be implemented during the second 18-month phase of the project.
- ***Scholarly Web Services on a Services Platform*** We will model existent digital humanities applications as web services and then deploy these or their proxies, and other web services, on robust platforms to help support many application developers in the humanities.
- ***Collections Interoperability*** We are selecting and recommending metadata standards for digital content that will allow for predictable use of important

scholarly material across all Project Bamboo Work Spaces and with all Bamboo tools. We are building an initial suite of adapters and connectors that facilitate easy, cross-collection use of these “target” collections within the Bamboo environment: content from the inter-university HathiTrust repository; selected content from the Perseus Digital LIbrary; and 400-years of English texts from the Text Creation Partnership (TCP). The CI group is collaborating with other Bamboo working groups and external experts to identify the most appropriate standards to implement for both content access and structural interoperability.

4. Interoperability and Shared Infrastructure for the Humanities

The BTP is establishing shared infrastructure, in the form of web services, interoperable collections, and organizational partnerships to meet the technology needs of humanities researchers and institutions.

- ***RESTful APIs*** Project Bamboo is modeling services as resources, and will leverage RESTful APIs to manage these resources across the Bamboo ecosystem.
- ***AuthN-AuthZ protocols*** Project Bamboo will delegate authentication to established institutional infrastructure, such as Shibboleth, and to social media identity providers. Ongoing work with Internet2 committees such as Shib-dev and COmanage will leverage edu- space experience and adopt new and existing software to meet Bamboo's requirements.

6. Future Work

Our next eighteen-month phase will primarily be concerned with implementing the Corpora Space Road Map, which will focus in terms of content on the 450 years of print culture in English from 1473 until 1923, along with the texts from the Classical world upon which that print culture is based. As an initial set of collections, we will be focusing on Perseus, TCP, ECCO, EEBO, HathiTrust, Google Books, Oxford Text Archive, and AUSTLit. Although we will begin with these corpora, we envisage a principle of user-driven growth. In gathering this material, we will move from the earlier to the later, because earlier texts pose more difficult textual and linguistic issues. In designing protocols that will address these issues of early texts, we believe we will anticipate most of the difficulties of the later material.

To enable users to work with these corpora, we will provide tools from the following five classes: 1. Curation; 2. Corpus Search and Discovery; 3. Collation; 4. Annotation; 5. Analysis/Visualization. We aim to build a modular environment that will attract content and collections from providers beyond the Bamboo partnership, and that will encourage the broad-based development of tools to operate on the data.

Notes

1. The partner institutions are Australian National University; Indiana University; Northwestern University; Tufts University; University of California, Berkeley; University of Chicago; University of Illinois, Urbana-Champaign; University of Maryland; University of Oxford; and University of Wisconsin, Madison
2. Theme Groups may be found here

Wandering Jew's Chronicle Research Archive

Cummings, James
James.Cummings@oucs.ox.ac.uk
University of Oxford, United Kingdom

Bergel, Giles
Giles.Bergel@merton.ox.ac.uk
University of Oxford, United Kingdom

1. Introduction

This poster will describe and demonstrate work done in creation of an online archive for research into the Wandering Jew's Chronicle (WJC). The WJC is a printed ballad published between 1634 and circa 1820 which survives in 22 known copies of 15 editions. These are held in ten libraries in Britain and the USA. The ballad itself outlines the succession to the throne of England from William I to a variable contemporary monarch depending on its date of publication. More specifically these are from the reign of Charles I until that of George IV, taking in seven monarchs in continuations from a core text. The succession of these monarchs is narrated by the supposedly immortal Wandering Jew of European legend. There is immense scholarly interest not only in the subject matter, but textually in the pattern of variations, the length and breadth of its publication and distribution. For a digital humanities perspective the textual history and relationships pose interesting problems for collation and textual analysis. Each one of the editions inherits a basic core text: some of these editions incorporate common continuations or variations, while others are textually idiosyncratic. The editions of the WJC are not only textually but also graphically interesting as most of the editions are illustrated with woodcuts of the monarchs described, and while some editions share woodcuts in common others employ copies or individual illustrations. Some of these editions are historically linked to others through relations within the book trade, while others are unauthorised or independent printings. The poster and demonstration will introduce the benefits of having gathered all the material relating to the WJC in a single place while demonstrating the technologies used to create the research archive.

2. Wandering Jew's Chronicle Research Archive

Surviving copies of the WJC are scattered: variously held in the Bodleian Library; the British Library; Cambridge University Library; Magdalen College Cambridge, Pepys Library; The University of Texas at Austin Harry Ransom Library; and the Brown University Library. The WJC research archive has created a digital archive in which surviving editions are united under a single authoritative citation and represented by:

- archivalquality images
- transcriptions markedup in TEI P5 XML
- visualization tools for comparing variations between texts and images
- bibliographical metadata
- scholarly commentary

It is hoped that by providing all of this in one location research into WJC can flourish in new and interesting ways. This resource helps to foster digital humanities research through tracing and expressing bibliographical, textual and iconological relations across a corpus of copies, variant editions, and versions of ballad texts, including their images and tunes. It is a valuable resource for those researching textual genealogy in the earlymodern period. It will impact the research of scholars of folklore, balladry, historiography, book history and textual studies.

The WJC research archive is an interdisciplinary project that touches on literary and textual studies; art history; historiography; popular culture and folklore; music and of course the digital humanities. It is hoped that in addition to its collaboration with the existing Bodleian Broadside Ballad Database, it will also collaborate with similar projects for creating digital archives of other ballads and related texts in a single coherent manner.

One of the key research objectives in the creation of the WJC research archive is the development of a convenient visualization tool for the diachronic representation and scholarly analysis of a popular representation of kingship and English historical continuity. The strengths and limitations for this purpose of the printed ballad form are of particular interest. The archive, and associated research, show both continuities and departures from at least three other comparable forms: oral kinglists; manuscript genealogical chronicles; and iconographic picture-galleries. From its first publication (c.1634), The WJC incorporated uptodate antiquarian research into the

legal documentary sources of royal legitimacy. The text may also be framed within the comparative progress of philological and folkloric inquiry into the origins of national culture and tradition over its span of publication. The WJC research archive seeks to embed this largely forgotten text within a larger history – of historical writing and memorialising in oral, literate and visual cultures, between the antiquarianism of the early seventeenth and the romanticism of the early nineteenth centuries.

3. Project Outcomes

The main outcome of this project to has been the creation of an archive of the WJC texttradition that facilitates the analytical comparison of its text and images. In one mode, the archive operates as a paralleltext edition, the usual way of representing a chronicle tradition in print, but greatly enhanced by the flexible display options afforded by digital publication. In particular this allows users to compare arbitrarilyselected transcriptions, revealing accidental, typographic and orthographic variation alongside more substantive literary insertions, emendations and continuations, including those supplied for ideological purposes. In another mode, the archive provides visualizations of surrogate images of original documents. 2 Certain editions are bibliographically related through historical relationships within the book trade, which is graphically revealed through publishers' imprints and their use of variant states or copies of the same wooden printingblocks. These relationships will be stated and made visible to the user, through a visualisation of the historical sequence of woodcut images, keyed to corresponding portions of the text.

A second but important outcome has been then provision of a complete citation of all known versions, editions and copies of the WJC tradition, creating common cataloguing standards for ballad texts in both broadside and chapbook formats that date from across the period. By drawing from materials held in various libraries, the project is promoting the importance of common cataloguing standards, scholarly collaboration and technical interoperability. The value of both bibliographically precise and semantically discerning cataloguing practices is demonstrated, at a time when the considerable promise of mass digitization is in danger of eclipsing the increasingly essential function of the authoritative scholarly catalogue. The WJC research archive engages with union catalogues such as OCLC WorldCat; the English ShortTitle Catalogue; the NineteenthCentury ShortTitle Catalogue; and is promoting the idea of a common catalogue of British ballads and chapbooks.

4. Conclusion

While the majority of the technologies used in the creation of the WJC research archive are mature and stable applications, in other aspects it has had to be more innovative in the application and combination of new technologies and bespoke programming. The use of lesstraditional digital humanities techniques will be demonstrated and explained alongside the poster.

It is hoped that in time the archive will be the centre-piece of a number of related projects about WJC. These include: a study applying imagerecogntion techniques to woodcut illustrations, in partnership with Professor Andrew Zisserman (Oxford Visual Geometry Group); an analysis of the registration and publication history of broadside ballads; an exploration of means of visualizing genealogical structure via earlymodern letterpress technology, in partnership with Mr. Paul Nash (Bodleian Library); research in the textual transmission of the ballad; and general researches in the genealogical and historical culture of the handpress period (c.14501820). It is hoped that these projects based on the WJC research archive will bring together scholars of balladry, book history, textual studies and digital humanities, and impact wider audiences interested in history, genealogy, folklore and popular culture.

Synergies: On the Production of a Sustainable, Open, e-Publication Infrastructure for the Academy

Eberle-Sinatra, Michael
michael.eberle.sinatra@umontreal.ca
Université de Montréal

This poster will present an overview of *Synergies: The Canadian Information Network for Research in the Social Sciences and Humanities*, a project funded by the Canada Foundation for Innovation to the order of $13 millions in their 2007 program 'Knowledge Management Resources for the Human and Social Sciences'. The poster will also offer a progress report on the technical challenges for harvesting and displaying various kind of information and data from across the 22 universities involved in this large scale infrastructure project as the project enters its last year of funding. It will be accompanied by a live demonstration of the web-based search interface launched in the spring of 2010.

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* not only brings that research into the mainstream of worldwide research discourse but also it legitimizes online publication in social sciences and humanities. The acceptance of this medium opens 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 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 website 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 ntegrates 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 is introduced within the broader context of scholarly communications. The Ontario region is 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.)

Stylometry with R

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

Rybicki, Jan
jkrybicki@gmail.com
Pedagogical University of Kraków, Poland

Stylometric studies, in all their variety of material and method, share two common features: the electronic texts they study have to be coaxed somehow to yield numbers, and the numbers themselves have to be processed with statistical software. Sometimes, the two actions are two independent parts of a given study. To give the simplest example, one piece of software is used solely to compile word frequency lists; then, one of the many commercial statistics packages takes over to extract meaning from this mass of words, draw graphs etc.

This approach works if standard statistical procedures are applied to the data. For multivariate word-frequency-based authorship attribution, for instance, once free AntConc produces word lists with frequencies, expensive (yet often university-licensed) SAS, SPSS or Statistica can be used to produce relative frequencies and correlation matrices thereof, and one of their many modules can in turn produce Principal Components (PCA), Cluster (CA), or Multidimensional Scaling (MDS) analyses. This is already a very comfortable situation when compared to what Burrows had to rely on in his seminal work on Jane Austen's style (Burrows 1987).

Yet, as stylometrists have begun to produce statistical methods of their own – to name but a few, Burrows's Delta, Zeta and Iota (Burrows 2002, 2006) and their modifications by other scholars (Argamon 2008, Craig and Kinney 2009, Hoover 2004a, 2004b) – commercial software, despite its wide array of accessible methods, becomes something of a straitjacket. This is why a number of dedicated stylometric solutions have appeared, targeting the specific analyses frequently used in this community.

Hoover's Delta, Zeta and Iota Excel spreadsheets are a pioneering and excellent example of this approach (Hoover 2004b). Constantly developed since at least 2004 (when one of the authors of this presentation received a CD-ROM with an early version), they have at least two major assets: they do exactly what the

stylometrist wants (with several optional procedures) and they only require spreadsheet software that has become (for better or worse) the standard on most computers in the world today. This has been especially helpful for uses in specialist workshops and classrooms; the student only needs additional (and, often, free) software to produce word frequency lists and he/she is ready to go. Yet Excel imposes one limitation: it is overkill in terms of memory usage, which results in the slowness of its processing. Also, the two-stage nature of the process (a separate piece of software prepares word lists that can be later automatically imported into the spreadsheet) might be something of a problem simply because it would take an experienced Visual Basic programmer to make Excel compile the word lists themselves.

In this respect, Juola's JGAAP can directly import texts in a variety of formats and perform a whole variety of authorship attribution tasks with an imposing variety of methods, statistical approaches and material on which they are based (Juola *et al.* 2006, 2008). These can be further expanded by experienced programmers in Java.

Java is also the language of another software solution which takes possibly an even broader approach. Craig's Intelligent Archive (in its many flavours), apart from performing certain stylometric tasks, is also a corpus organizer; once the initial work of registering texts is done, it allows a versatile combination of individual texts and groups of texts (Craig and Kinney, 2010).

A new trend in producing stylometric tools is associated with R, a GNU project, a language and environment for statistical computing and graphics (www.r-project.org) in the image of its commercial counterpart, S. While it came into its version 1.0 in 2000, it has been used for analyses on language only recently. Its strongest promoters in this community include authors of corpus-linguistics-oriented books on the usage of R, Baayen (2008) and Gries (2009). R has already been used in authorship attribution research by Jockers *et al.* (2008, 2010).

The scripts presented in this poster have begun with the first author's participation in an R workshop taught by Gries at University of Leipzig's 1st European Summer School "Culture and Technology." Very soon, a series of R scripts appeared, targeted at a variety of experiments with Delta and other distance measures. Very soon, too, it became evident that R is capable of processing texts and statistics in a fraction of the time needed by other tools. Also, R scripts can be relied on for doing the whole work themselves, from manipulating texts (typically, to produce word lists) all the way to graphing the results (and that in a great variety).

These capabilities were put tu use in the study of Delta's dependence on studied texts' sizes (Eder 2010), and on the behaviour of Delta at a variety of intervals in the word frequency rank lists in a variety of languages (Rybicki and Eder 2011). The R environment permitted us to cover huge statistics at unprecedented rates (often thousands of Delta iterations per hour in corpora of a hundred full-size novels). It also permitted us to use other statistical methods, such as PCA, CA or MDS, or those rarely used so far in this field – such as bootstrap consensus trees based on a study of Papuan languages by Dunn *et al.* (2005, quoted in Baayen 2008: 143-47). What is more, the entire analysis – from loading texts to final results in numeric and graphic form – can be accomplished with a single script. Our Delta script, for instance, did all the work: it processed electronic texts to create a list of all the words used in all texts studied, with their frequencies in the individual texts; normalized the frequencies with z-scores (if applicable); selected words from stated frequency ranges for analysis; performed additional procedures that (usually) improve attribution, such as Hoover's automatic deletion of personal pronouns and culling (automatic removal of words too characteristic for individual texts); compared the results for individual texts; performed a variety of multivariate analyses; presented the similarities/ distances obtained in tree diagrams; finally, produced the above-mentioned consensus tree – a new graph that combined many tree diagrams for a variety of parameter values (Fig. 1). It was our aim to develop a general platform for multi-iteration attribution tests; for instance, an alternate script produced heatmaps to show the degree of Delta's success in attribution at various intervals of the word frequency ranking list (Fig. 2).

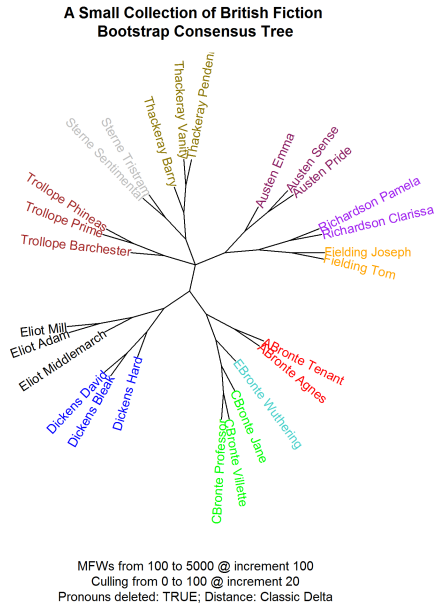


Figure 1. An example of a consensus tree diagram generated by R

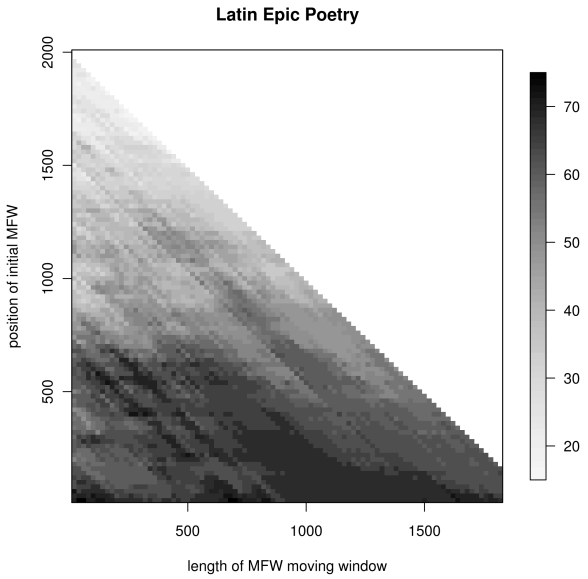


Figure 2. An example of a heatmap generated by R, showing the percentage of successful Delta authorship attributions at combinations of word list length and initial word rank in frequency lists

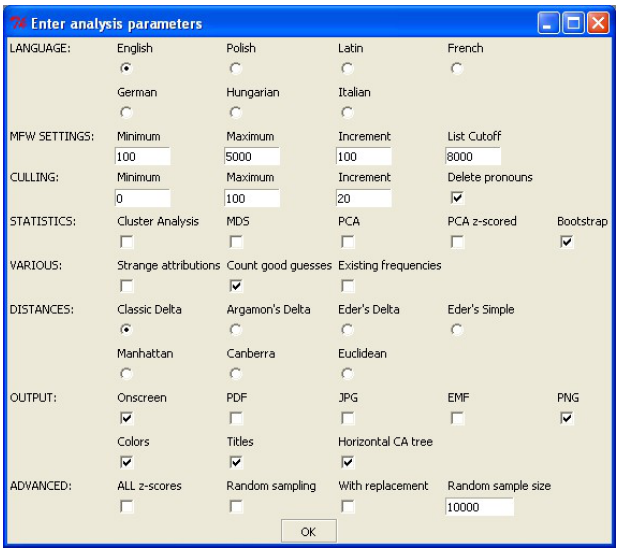


Figure 3. The GUI for the R script used to generate the diagram in Fig. 1

What is more, despite the fact that the R environment might daunt (digital) humanists with its initially steep learning curve, it soon became evident that ready-made tools can be developed to make the full power of R accessible even to inexperienced users with its capability of working with Tcl/Tk graphic user interfaces (Fig. 3) prepared by the second author for two seminar groups of his MA students, who were asked to make the switch to R from the more traditional software, with success: one of these groups has already successfully defended their completed theses on authorship attribution and stylistic variety in a good number of corpora.

References

Argamon, S. (2008). 'Interpreting Burrows's Delta: Geometric and Probabilistic Foundations'. *Literary and Linguistic Computing.* 23(2): 131-47.

Baayen, H. R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R.* Cambridge: Cambridge University Press.

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(3): 267-87.

Burrows, J. F. (2006). 'All the Way Through: Testing for Authorship in Different Frequency Strata'. *Literary and Linguistic Computing.* 22(1): 27-48.

Craig, H., Kinney, A. F. (eds.) (2009). *Shakespeare, Computers, and the Mystery of Authorship.* Cambridge: Cambridge University Press.

Craig, H., Whipp, R. (2010). 'Old spellings, new methods: automated procedures for indeterminate linguistic data'. *Literary and Linguistic Computing.* 25(1): 37-52.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A., Levinson, S.C. (2005). 'Structural Phylogenetics and the Reconstruction of Ancient Language History'. *Science.* 309: 2072-75.

Eder. M. (2010). 'Does Size Matter? Authorship Attribution, Small Samples, Big Problem'. *Digital Humanities 2010: Conference Abstracts.* King's College London, 7-10 July 2010, pp. 132-34.

Gries, S. Th. (2009). *Statistics for Linguistics with R: A Practical Introduction.* BerlinNew York: Mouton de Gruyter.

Hoover, D. L. (2004a). 'Testing Burrows's Delta'. *Literary and Linguistic Computing.* 19(4): 453-75.

Hoover, D. L. (2004b). 'Delta Prime?'. *Literary and Linguistic Computing.* 19(4): 477-95.

Jockers, M. L., Witten, D. M., Criddle, C. S. (2008). 'Reassessing Authorship in the 'Book of Mormon' Using Delta and Nearest Shrunken Centroid Classification'. *Literary and Linguistic Computing.* 23(4): 465-91.

Jockers, M. L., Witten, D. M. (2010). 'A Comparative Study of Machine Learning Methods for Authorship Attribution'. *Literary and Linguistic Computing.* 25(2): 215-23.

Juola, P., Noecker, J., Ryan, M., Zhao, M. (2008). 'JGAAP3.0 – Authorship Attribution for the Rest of Us'. *Digital Humanities 2008: Book of Abstracts.* University of Oulu, 25-29 June 2008, pp. 250-51.

Juola, P., Sofko, J., Brennan, P. (2006). 'A Prototype for Authorship Attribution Studies'. *Literary and Linguistic Computing.* 21(2): 169-78.

Rybicki, J., Eder, M. (2011). 'Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?'. *Literary and Linguistic Computing.* 26: (forthcoming).

Pleiades: an un-GIS for Ancient Geography

Elliott, Tom
tom.elliott@nyu.edu
Institute for the Study of the Ancient World, New York University

Gillies, Sean
sean.gillies@gmail.com
Institute for the Study of the Ancient World, New York University

---

Pleiades (http://pleiades.stoa.org) is an open-access digital gazetteer for ancient history. It has recently been described in an important report from the Council on Library Information Resources as a "prominent digital project ... within the realm of classical geography." (Babeau 2010) It provides stable Uniform Resource Identifiers (URIs) and helpful Atom, HTML, JSON, and KML representations for tens of thousands (and growing) of geographic entities. It invites scholars, students, and enthusiasts worldwide freely to use, create, and share historical-geographic information related to these entities via its web application. Built on the Classical Atlas Project (1988-2000), which produced the Barrington Atlas of the Greek and Roman World, Pleiades is co-organized by the Institute for the Study of the Ancient World (NYU) and the Ancient World Mapping Center (UNC Chapel Hill). With fresh funding from the National Endowment for the Humanities (2010-2013), Pleiades is solving problems in information design, making more data accessible, and increasing interoperability with other web-based resources treating the geographical, textual, visual and physical culture of antiquity. Our poster, to be accompanied by hands-on software demonstrations, will highlight new and ongoing collaborations and summarize the findings since the publication of our paper in Digital Humanities Quarterly in early 2009 (Elliott and Gillies 2009) and Elliott's general presentation of the project at DH2009 which we believe are of significant interest to digital humanities practitioners and scholars across sub-disciplines.

The poster will foreground a unique aspect of the Pleiades effort: its "un-GIS" approach to historical geography. Where conventional Geographic Information System (GIS) data models privilege geometry – requiring a point, line or polygon with which to associate such "attribute data" as toponyms, time periods and the like – Pleiades embraces the inevitable

sparseness, ambiguity, and contingency of historical knowledge. Pleiades models historical geography as a graph of relationships between conceptual places/ spaces, names, locations, and time periods rather than as layered views of tables containing measured locations with associated descriptive data fields. We find that our approach opens up a range of flexible capabilities not afforded by a traditional GIS approach, including:

- Identification and representation of geographic features that have no known locations, or that can be located only vaguely, roughly, or in relationship to each other
- Representation of the connectedness of features
- Change of location or properties of a feature over time
- Aggregation of temporally varying features into conceptual places or spaces that reflect ancient practice or modern scholarly method
- Stable identifiers for features, places and names that can be addressed directly by other web applications and cited reliably by students and scholars

The features listed above don't come for free. They bring extra complexity, and this additional complexity might be considered a weakness of the un-GIS approach (or not, as we'd like discuss with poster session attendees):

- Vocabularies for describing vague or rough locations are immature
- Algorithms for reasoning on vague or rough locations are immature or proprietary
- Relative locations need closure at some level if analysis is to be permitted, and there is uncertainty about where to stop
- Vocabularies for describing connectedness are immature
- Time adds both technical and social complexity; event-based conceptual models (like that proposed by Mostern and Johnson (2008) embracing the inseparability of time and space may be required

Pleiades' web-oriented approach – especially its emphasis on stable URIs and multiple standard formats for content serialization – opens up a wide range of interoperability options. Current collaborations in this domain will be highlighted, including:

- Ancient World Image Bank: http://www.nyu.edu/awib/
- Digital Atlas of Roman and Medieval Civilization: http://darmc.harvard.edu/
- Google Ancient Places: http://www.ecs.soton.ac.uk/about/news/332
- Epigraphische Datenbank Heidelberg: http://www.uni-heidelberg.de/institute/sonst/adw/edh/
- The Inscriptions of Roman Cyrenaica: http://ircyr.kcl.ac.uk/
- Open Encyclopedia of Classical Sites: cf. in http://googleblog.blogspot.com/2010/07/our-commitment-to-digital-humanities.html
- Nomisma.org: http://nomisma.org/
- Papyri.info: http://papyri.info
- The Portable Antiquities Scheme: http://finds.org.uk/
- PELAGIOS: http://pelagios-project.blogspot.com/

At a moment when interest in Humanities GIS is growing (Rumsey 2009), this poster is especially germane. Critical engagement with the methodological and theoretical facilities and inadequacies of GIS is essential if we are fully to integrate spatial approaches into the humanist's toolkit and evaluate the products of their use in research, teaching and outreach. The perspective of the Pleiades team is firmly rooted in practice, arising from extended, hands-on work with a large, complex collection of historical information that exemplifies many of the challenges and opportunities humanists face daily, as well as hard-earned experience in conventional GIS projects. We believe that our critique of conventional GIS, and the alternative approach that has arisen from it, can inform the design of other humanities projects and provoke further innovation. Moreover, as the realization of the Pleiades model has been predicated upon an embrace of web-based interoperability and "crowd-sourced" content curation, this poster speaks directly to the DH2010 conference theme of "big-tent digital humanities."

References

Babeau, A. (2010). *Rome Wasn't Digitized in a Day: Building a Cyberinfrastructure for Digital Classicists.* `http://www.clir.org/activities/details/infrastructure.html`.

Elliott, T. and Gillies, S. (2009). 'Digital Geography and the Classics'. *Digital Humanities Quarterly.* 3. `http://digitalhumanities.org/dhq/vol/003/1/000031/000031.html`.

Mostern, R. and Johnson, I. (2008). 'From named place to naming event: creating gazetteers for history'. *International Journal of Geographical Information Science.* 1091-1108.

Rumsey, A. S. (2009). *Summary and Report of the UVA Scholarly Communications Institute 7: Spatial Technology and the Humanities, June 28-30 2009.* http://www.uvasci.org/archive/spatial-technologies-and-methodologies-2009.

Visualizing Sound as Functional N-Grams in Homeric Greek Poetry

Forstall, Christopher
forstall@buffalo.edu
State University of New York at Buffalo

Scheirer, Walter J.
wjs3@vast.uccs.edu
University of Colorado at Colorado Springs

The question of the stylistic integrity of Homer's corpus is a venerable one. For centuries, diverse models, subjective as well as quantitative, have claimed to explain the composition of the *Iliad* and *Odyssey*: some scholars have seen the poems as the work of a single, literate genius (West, 2001, 3); others as a collective multitext, the superposition of generations of continually-changing performances handed down from one illiterate bard to the next (Nagy, 1996, 107ff.). Often much of the support for these claims is the perceived homo- or heterogeneity of the text. What is at stake in these examinations is larger than a nineteenth-century romantic notion of the artist and his genius; recent studies have used the structure of the ancient Greek epics to examine how cognition structures spoken poetry, and how the sounds of poetry in turn give structure to our thought (Peabody, 1975, 168ff.). A connection between low-level phonetic patterns and larger-scale poetics is not unique to oral composition, but has been shown to be equally active in literate authors as well (Brierley and Atwell, 2010).

The work in progress presented here attempts to examine internal heterogeneity in Homer's language using the tools of computer-based authorship analysis. As stylometric tools become finer-grained, scholars such as Hoover (2007) and Andreev (2010) have turned their gaze from the characterization of an author or corpus as a whole to considerations of an author's stylistic evolution over time, and the differences between and even within individual works. Previously, the authors of this poster have used character- and word-level functional n-grams to compare Homer's two epic poems to one another and to later written text (Forstall and Scheirer, 2010a). We have also adapted the functional n-gram to metrical data (Forstall, Jacobson and Scheirer, 2010). In the present research, we attempt to characterize the internal sound structure

of Homer's epics using functional n-grams at the word, character and metrical levels.

We ask,

- How homogeneous is the author signal within a large work?
- Do the areas frequently identified as later additions stand out?
- Can internal patterns help us understand a poem's composition?

While statistical studies of Homer have been made before, it is often difficult for the critic to move comfortably between the numbers and the subjective experience of interpreting poetry. David W. Packard, in pioneering computational work on sound patterns in Homer, cautioned that "we cannot expect to identify expressive passages merely by counting letters" (Packard, 1974). More recently, Marjorie Perloff, noting that the significance of sound is all too often overlooked in poetry criticism, has laid part of the blame on "'scientific' prosodic analysis," which "has relied on an empiricist model that allows for little generalization about poetic modes and values: the more thorough the description of a given poem's rhythmic metrical units, its repetition of vowels and consonants, its pitch contours, the less we may be able to discern the larger contours of a given poet's particular practice, much less a period style or cultural construct." (Perloff and Dworkin, 2009, 2) In this poster we focus on visualizing the data in ways that bridge the gap between empirical data and the subjective experience of interpreting poetry. We take our inspiration from work such as that of Plamondon (2009) and the online *Poetess Archive*, which has shown that innovation in how we visualize data is vital to connecting computing with humanities scholarship. In particular, Plamondon used color to represent multi-parameter sound data over individual poems, allowing a subjective appreciation of the poem's structure based on objective values at a glance.

We divide the poem into samples of various sizes, and calculate n-gram frequencies for the most common features. We then use principal components analysis to concentrate the variance among fewer variables. The top three principal components are assigned to three component color channels: red = PC1, green = PC2, blue = PC3. Each sample is visualized as a color which simultaneously represents three parameters, each potentially comprehending the most important aspects of a much larger feature set. The flow of sound in the poem may be seen as a gradient with local and large-scale variation (see Figure 1). As a control, we also treat a text of Homer's poetry in which the order of the lines has been randomized. This is in part a response to the sobering results shown by Eder (2010), who made a strong case that authorship analysis was unreliable at samples fewer than several thousand words, and was improved with randomization in sampling. It may be that smaller samples are less reliable at the author level precisely because they are sensitive to internal patterns in the text, which the randomization should smooth over.

The color gradient produced by this visualization of PCA is useful to the classical philologist precisely because of its subjective quality; yet a more definitive analysis of the epics' internal heterogeneity is also desirable. Which sections are the "most different"? Are units which are functionally related, for example the type scenes which are played out over and over by different characters (Lord, 2000, 68 ff.), more consistent than the poem as a whole? Is the difference between the *Iliad* and *Odyssey* greater than the variation within each poem? In addressing such questions, the classicist cannot help but be biased: certain features, certain passages take on prominence at the expense of others. Here we turn to unsupervised classification for an answer which encompasses all of the text at once, and has no literary bias. Using the same features as for PCA, we now perform k-means classification for various numbers of categories; again, the randomized text is used as a control. In forcing the algorithm to subdivide our sample set into an arbitrary number of classes, we make no specific assumptions about the structure of the poem. Instead we ask, how is variation distributed within the text? For example, with only two or three classes, we find that samples from all are relatively evenly distributed within and between the two poems. With more classes, we find that some tend to be found more in one poem or the other in the ordered version of the text, while in the randomized text samples from all classes are found throughout. The computer-assigned, discrete classifications may be arbitrarily assigned to colors, which are then displayed alongside the continuously varying PCA data to contrast the more subjective, human-interpreted view of the poem's heterogeneity with the entirely objective computer-based analysis.

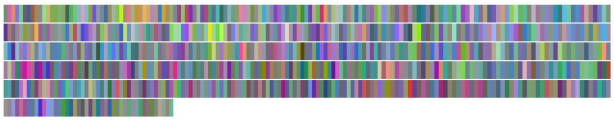


Figure 1

References

Andreev, V. S. (2009). 'Patterns in Style Evolution of Poets'. *Digital Humanities 2009.* University of Maryland, June, 2009, pp. 52–53.

Brierley, C., Atwell, E. (2010). 'Holy Smoke: Vocalic Precursors of Phrase Breaks in Milton's *Paradise Lost*'. *Literary and Linguistic Computing.* 25(2): 137–151.

Eder, M. (2010). 'Does Size Matter? Authorship Attribution, Small Samples, Big Problem'. *Digital Humanities 2010.* King's College London, July 2010, pp. 132–133.

Forstall, C. W., Scheirer, W. J. (2010). 'Features From Frequency: Authorship and Stylistic Analysis Using Repetitive Sound'. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science.* .

Forstall, C. W., Jacobson, S. J., Scheirer, W. J. (2010). 'Evidence of Intertextuality: Investigating Paul the Deacon's *Angustae Vitae*'. *Digital Humanities 2010.* King's College London, July 2010, pp. 294–295.

Hoover, D. L. (2007). 'Corpus Stylistics, Stylometry, and the Styles of Henry James'. *Style.* 41(2): 160–189.

Lord, A. B. (2000). *The Singer of Tales.* Cambridge, MA: Harvard University Press.

Nagy, G. (1996). *Poetry as Performance: Homer and Beyond.* Cambridge, UK: Cambridge University Press.

Packard, D. W. (1974). 'Sound Patterns in Homer'. *Transactions of the American Philological Association.* 104: 239–260.

Peabody, B. (1975). *The Winged Word: A Study in the Technique of Ancient Greek Oral Composition as Seen Principally Through Hesiod's "Works and Days".* Albany, NY: State University of New York Press.

Perloff, M., Dworkin, C. (eds.) (2009). *The Sound of Poetry, the Poetry of Sound.* Chicago, IL: University of Chicago Press.

Plamondon, M. (2009). 'Computational Phonostylistics: Computing the Sounds of Poetry'. *Chicago Colloquium on Digital Humanities and Computer Science.* Illinois Institute of Technology, November 2009.

Mandell, L. (ed.). *The Poetess Archive.* `http://www.poetessarchive.com` (accessed 09/14/2010).

West, M. L. (ed.) (2001). *Studies in the Text and Transmission of the Iliad.* Munich: K. G. Saur.

DHAnswers: Building a Community-Based Q&A Board for the Digital Humanities

Gilbert, Joseph
joegilbert@virginia.edu
University of Virginia

Meloni, Julie
jcmeloni@gmail.com
University of Victoria

Nowviskie, Bethany
bethany@virginia.edu
University of Virginia

Sinclair, Stéfan
sgs@mcmaster.ca
McMaster University

1. Genesis and Motivation

In late September 2010, the Association for Computers and the Humanities (ACH), together with *ProfHacker*, a technology and productivity blog hosted by the *Chronicle of Higher Education*, announced the launch of “DHAnswers,” a community-based question-and-answer board, at http://digitalhumanities.org/answers/.

With DHAnswers, the Outreach Committee of the ACH sought to address both an opportunity and a problem we detected with existing communication venues for digital humanists, such as Twitter, the long-standing *Humanist* discussion list, and individual blogs. We identified a need for DHAnswers after observing the burgeoning and helpful “big tent” digital humanities conversation happening on Twitter —and the frequency with which answers to questions posed by members of that community exceeded Twitter's 140-character limit for tweets, or required near-impossible sharing of a code snippet. Other exchanges resulted in extended and hard-to capture conversational threads, generally lost in a matter of weeks as older tweets were purged from search engines. We also noted that many questions asked on Twitter were more specific than those generally asked on the Humanist discussion list, or were more basic than a newer member of the DH community might feel comfortable posing on specialist mailing lists for software or standards.

A small team from ACH and ProfHacker worked behind the scenes (with most software development and extension undertaken by the University of Virginia Library's Scholars' Lab and stemming from work on a Spatial Humanities gateway site [http://spatial.scholarslab.org]) to create a useful communication platform, with pre-defined topic categories (enriched by input from the Executive Council of the ACH) to help filter and focus discussion. In addition to a viable, open-source platform for discussion, however, this project needed people. We therefore recruited approximately 25 digital humanities colleagues from around the world, working in different disciplines and with differing areas of expertise to test and cultivate the system. We were mostly concerned with having a friendly group of people helping to pre-populate the site with sample questions and answers, who could be at the ready in the first months after release, to monitor the various notification features we had set up (RSS feeds, email options, and automatic Twitter messages) to ensure that questions were answered promptly and the proper communities were alerted to relevant discussions. We also asked these volunteers to help us keep the discourse on DHAnswers positive and friendly. Thanks to the efforts of this group, we were able to launch the site with a small amount of content present in each of the following categories:

- Applications, Tools, Formats
- Databases & Data Structures
- Interfaces, Design & Usability
- DH in the Classroom
- Markup & Metadata
- Programming
- New Media & Games
- Project Management & DH Professions
- and “About DHAnswers”

Reaction to the public release of DHAnswers was enthusiastic. Within a week, nearly 200 people had registered for accounts and created nearly 300 responses in the site's question-and answer threads. One month in, it is rare to see a question go unanswered for more than a few hours—and thanks to Twitter integration, many questions garner immediate response.

2. Technical Implementation

DHAnswers leverages an open-source bbPress platform, which employs PHP and MySQL and is related to the popular WordPress content management system. With a flexible system for creating new stylistic themes and adding new functionality through plugins, bbPress allowed us to create a custom set of features built upon a supported and extensible base architecture. To keep the site as lean and usable as possible, and in the hope of creating a self-explanatory service, we simplified or removed altogether a number of features from the out-of-the-box bbPress application.

Given a strong digital humanities presence on Twitter, integrating DHAnswers with that target community was deemed imperative to the site's success. We introduced purpose-built plugins to broadcast new questions from the @DHAnswers Twitter account, giving both DHAnswers members and non-members alike a real-time peek into the ongoing conversation. Initially, we enabled even deeper integration with Twitter: site members could tweet a message to the @DHAnswers account in order to create a new question on the DHAnswers site. In the end, we realized that the complexity of rules and procedures for this connection would require a lengthy, nuanced explanation to each user. The added Twitter functions thus ran counter our goal of a straightforward tool that scholars could easily integrate with their normal communication methods, and we removed them.

To facilitate fast-paced conversation, we created email notifications and RSS feeds for all questions and/ or for a selection of “favorite” topics. To reward regular user involvement, the Scholars’ Lab create a new plugin to add various “badges,” small medal-like symbols that indicate a certain number of posts made or questions answered, to user profiles. We feel certain that the robust notification system and subtle reward mechanism for constructive behavior on the site spurred rapid growth and a quick response time for new questions.

3. Response from the DH Community

The multiple broadcast methods (Twitter, RSS, email) for new questions and answers helped bring together what some feel are disparate groups within the digital humanities: those who are on Twitter and those who are not. Although Twitter integration is a key feature of DHAnswers, Twitter participation is not required of DHAnswers users. Instead of relying solely on the segment of the DH community on Twitter, and thereby narrowing our audience rather than expanding it, DHAnswers has focused on building its own community: the multiple broadcast methods, our "reward" badges mentioned previously, the ability for

users to select "favorite" posts within the system, and our administrative caretaking have all worked to create a community of sharing and mentoring. On more than one occasion, a new user has come to DHAnswers to ask a question, never having set a virtual foot within the initial Twitter community that inspired it (and never intending to), and has found himself or herself surrounded by senior members of both the DHAnswers community and the broader digital humanities community, ready to answer questions from all. Contrary to initial expectations, conversations on DHAnswers have centered on pedagogical and institutional questions, such as building a Digital Humanites center or designing a curriculum, rather than technical inquiries on specific processes. Also, unexpected disciplines, such as archaeology, have established lively running conversations whereas more explicitly digital fields—like media studies—have had relatively limited involvement.

One of the differences between user interactions in the DHAnswers forum and within other social networking spaces is that the inherent expectation of asynchronous responses allows for fruitful participation by more users and according to their own terms of time management. Instead of constantly filtering an incoming information stream via Twitter, in which a response more than a few hours later is viewed as out of date and nearly useless, DHAnswers participants understand that a threaded discussion will remain in plain sight for several days, thus providing the mental space for better discussion. When comments are added to the thread, the "freshness" of the thread—as well as the notifications in place throughout the system—continue to keep the information in view; this process extends the life of the question and, in turn, the visibility of the answers. In addition, we find that users are taking advantage of the ability to "tweet this question"—re-posting questions of interest to them on Twitter—and therefore are helping to ensure that questions get useful answers and important discussions remain in the public eye.

4. Analysis

In the spirit of ongoing attempts to define and describe the Digital Humanities community (see for instance Svensson's "The Landscape of Digital Humanities" in *Digital Humanities Quarterly* vol. 4 no. 1, Summer 2010), DHAnswers provides an interesting opportunity to gather insights about who digital humanists are and what they do, or at least some of who they are and some of what they do (see also Bethany Nowviskie's article "DH Answers by the Numbers" in the *Chronicle of Higher Education's ProfHacker* blog for December 8, 2010). The value of such insights is of course moderated by the specific circumstances of the site, an English-only resource sponsored by two predominantly North American organizations (ACH and ProfHacker), with strong links to an existing Twitter-based community.

In its first month of its existence (essentially October 2010), DHAnswers recorded over 19,000 pages viewed by some 5,575 visitors from 64 countries (these data have been collected by Google Analytics and only include traffic to the main site, not RSS and Twitter feeds). The geographical distribution indicates a predominance of visitors from the USA, but also reveals emerging digital humanities regions such as Australia and Japan.

5,575 visits came from 64 countries/territories

1.	United States	3,662	65.69%
2.	Canada	631	11.32%
3.	United Kingdom	362	6.49%
4.	France	193	3.46%
5.	Australia	157	2.82%
6.	Japan	77	1.38%
7.	Germany	63	1.13%
8.	Belgium	57	1.02%
9.	Austria	36	0.65%
10.	Italy	34	0.61%

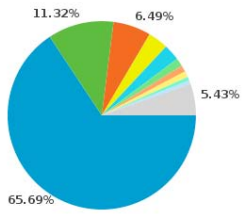


Figure 1

Similarly, the geographical distribution of site visits within the United States can be revealing of regional activity in digital humanities.

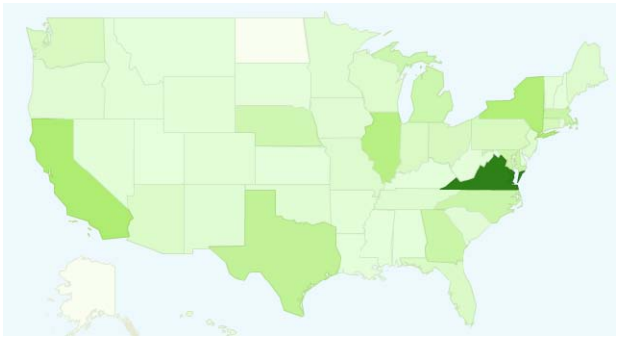


Figure 2

Also noteworthy is that in the first month over 250 users have registered to DHAnswers and there have been nearly 600 posts. The most frequently visited page, after the home page, is a topic on defining digital humanities, which further reinforces the community's desire to understand itself.

1.	/answers/	6,413
3.	/answers/topic/what-s-digital-humanities	481
5.	/answers/topic/doing-dh-v-theorizing-dh	290
6.	/answers/topic/what-are-some-useful-tools-for-creating-timelines	259
7.	/answers/forum/applications-tools-formats	253
8.	/answers/topic/how-co-we-introduce-undergraduates-to-the-digital-hu	238
9.	/answers/topic/visualization-in-digital-humanities-what-are-the-possibi	235
10.	/answers/topic/whats-in-your-applications-folder	211
11.	/answers/topic/help-us-design-a-dh-workspace	202
12.	/answers/topic/best-dh-project-management-system	191

Figure 3

In addition to further analysis of the web logs, we will spend some time analyzing and interpreting the actual content of the DHAnswers posts. For instance, the graph below, generated by Voyeur Tools, shows that the word “like” was the third most common content word in all of the posts, and fairly consistently present across the corpus (the trend column, which indicates relative frequency by topic). The word carries several meanings, of course, but a closer examination of the usages suggests a notable predilection of digital humanists to express preferences and make comparisons.

Words in the Entire Corpus

Frequencies	Count	Trend
digital	260	
http	256	
like	213	
i'm	212	
post	183	
humanities	172	
use	171	
project	157	

Figure 4

References

Meloni, Julie (2010). 'Announcing Digital Humanities Questions & Answers (@DHAnswers)'. *ProfHacker: The Chronicle of Higher Education.* . http://chronicle.com/blogs/profhacker/announcing-digital-humanities-questions-answersdhanswers/26544.

Nowviskie, Bethany (2010). 'DH Answers by the Numbers'. *ProfHacker: The Chronicle of Higher Education.* . http://chronicle.com/blogs/profhacker/dh-answers-by-the-numbers/29307.

Svensson, Patrik (2010). 'The Landscape of Digital Humanities'. *Digital Humanities Quarterly.* 1. http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

Pedagogy & Play: Revising Learning through Digital Humanities

Harris, Katherine D.
katherine.harris@sjsu.edu
Department of English, San Jose State University

In Digital Humanities circles, we often talk about collaboration between disciplines, among scholars, and with technologists. While progress in the field is nurtured certainly by this type of research, what of our students? How are we shepherding Digital Humanities to those undergraduates who could most benefit from exposure to collaborative tools or humanities computing strategies? Happily, HASTAC has been addressing pedagogy, most specifically with Cathy Davidson's post "Research is Teaching" (http://www.hastac.org/blogs/cathydavidson/research-teaching) and the wildly successful forum "Teaching with Technology and Curiosity" (http://www.hastac.org/forums/hastac-scholars-discussions/teaching-technologies). As is evident from the Digital Humanities Zotero group (http://www.zotero.org/groups/30) there are some relevant, engaging courses being taught in and around Digital Humanities. But, how can Digital Humanities engage with, even alter, traditional pedagogy?

Collaboration, shared knowledge, open access, extra-disciplinarity. These are the major tenets of Digital Humanities. However, what is missing in this list is something required of all digital projects: play. Roger Caillois qualifies this type of unstructured activity as “an occasion of pure waste: waste of time, energy, ingenuity, skill” (*Man, Play and Games* 2001; 6). This lack of structure, leads to exploration, discovery, and production of knowledge in ways that were only imagined twenty years ago. Typically though we don't allow our students this sense of play in their traditional studies. Especially in literary studies, we supply students with the end-product but don't expose them to the theories and the methodologies always. We separate those kinds of issues into other courses (e.g., Introduction to Literary Criticism or Introduction to Research Methods). When faculty bring a particular perspective, for example textual studies or feminist theory, to a classroom setting, the methods for exploring and discovering aren't exposed to students. Instead, we're offering them the one big major tool, close reading, for their arsenal.

Students then live with some anxiety that there's one way to read a text and, more often, ask "how does the professor want me to read this?" It becomes a guessing “game” instead of an exploration and discovery of the literature. In the final essay, we expect students to offer a discovery, a research paper, or an analysis. But, if we haven't exposed them to the methodology and the theory, how can they adequately achieve a true exploration of the literature? In this way, the course becomes a game with an outcome, consequences, and rigid rules. Using Digital Humanities strategies, I want to instill a sense, even if it's artificial, that literary studies are a “free and voluntary activity, a source of joy and amusement” as Caillois defines “play” (6).

With this poster, I will explore invigorating an undergraduate education with a sense of play by specifically incorporating the major tenets of Digital Humanities and Caillois' typology of play:

1. *Free*: in which playing is not obligatory; if it were, it would at once lose its attractive and joyous quality as diversion;
2. *Separate*: circumscribed within limits of space and time, defined and fixed in advance;
3. *Uncertain*: the course of which cannot be determined, nor the result attained beforehand, and some latitude for innovations being left to the player's initiative;
4. *Unproductive*: creating neither goods, nor wealth, nor new elements of any kind; and, except for the exchange of property among the players, ending in a situation identical to that prevailing at the beginning of the game;
5. *Governed by rules*: under conventions that suspend ordinary laws, and for the moment establish new legislation, which alone counts;
6. *Make-believe*: accompanied by a special awareness of a second reality or of a free unreality, as against real life. (9-10)

Before being able to articulate this type of change in undergraduate curriculum, I had to understand it myself. I attended ThatCamp Bay Area in October 2010 to gain some understanding, and, well, to feel uncomfortable. I wanted to immerse myself in areas that were not so familiar to me, sessions where I couldn’t be an authority. The invigorating aspect to the two days’ of sessions was that no matter how hard I tried to avoid familiar topics, I found myself

reflecting on the intersection between my work and all of the cool, interesting work being discussed. Steve Ramsay (http://lenz.unl.edu/wordpress/?p=266) was right; I was prepared to be the dumbest person in the room, and that prepared me for being inspired by the cross/multi/extradisciplinary work that so many people came together to discuss.

Big questions plagued me during and after sessions, even at the bootcamps. But, these weren't questions of despair; rather they were invigorating because they required that I think about pedagogy and curriculum in a different fashion, but they were there nonetheless: How can this apply in the classroom? How can I teach my students some of this technology without sacrificing content? Is this the content then in a Digital Humanities course? What kind of Humanistic inquiry comes from integrating tools with literary studies? How can I educate my colleagues about Digital Humanities using geo-referencing as an example? How can GIS impact my work on history of the book. But mostly I just wanted to play with all of the toys in order to explore what kind of Humanistic inquiry is possible. I wanted to see what happened when a major corpus of work was available; what questions could I come up with, because I don't have any to start with. Perhaps if I had a chance to play, though, I could find something.

In one session, Linguist Adita Muralidharan lead the group through algorithms that could parse n-grams in large datasets. The question came up: How do you know what questions to ask of such large amounts of data? For instance, Franco Moretti and Matthew Jockers' use of massive quantities of 19th-century novels to collate developments in the genre (http://www.thevalve.org/go/valve/article/from_the_che_text_mining_and_data_digging_as_the_ humanities_go_google/). While neither project can search for metaphor, irony, and humor, I can only guess what kinds of extrapolations could come out of just fooling around with the data, searches and results. But, I wouldn't know unless I got my hands on the tools and the dataset. Can I be allowed to do research when I don't know what the question will be, let alone the answers? Playfulness, see?

And this is the crux of the entire weekend – playfulness and imagination is perhaps something that academics and scholars have moved away from, something that is stolen from us as we move into full time positions. And we've done this to our students in a way.

Over the last two years, I have begun focusing my teaching on incorporating digital tools into my undergraduate classroom. This is often a nuanced decision made in heavy consultation with an Instructional Designer. Now, I teach three kinds of courses that interact at some level with digital tools, Digital Humanities and the typology of play:

1. TechnoRomanticism (http://www.sjsu.edu/faculty/harris/TechnoRom_F09/News.htm): We create our own digital edition of Mary Shelley's Frankenstein. Along the way, we create a collaborative timeline using MIT's SIMILE & Timeline script (http://www.simile-widgets.org/timeline/). We don't even begin to create a website until some of the preliminary assignments are done -- assignments that look at the construction of this novel, both linguistically and bibliographically. Every 2 weeks, we held a workshop on some digital assignment and acquired 1 new skill, not even necessarily a new tool, but a skill. Even those not accustomed to posting to forums and blogs got something out of it.
2. Digital Humanities: The Death of Print Culture? (http://www.sjsu.edu/faculty/harris/DigLit_F10/Introductions.htm) I'm teaching this one right now, and we're theorizing all facets of Digital Humanities while at the same time critiquing the tools for our thinking and dissemination. We're in Week 11, and now they're really seeing the benefits and pitfalls of Digital Humanities. We will also explore multi-modal arguments, i.e., the video essay.
3. And a third type of course, one in which content and Digital Humanities are intertwined -- the British Literature survey course 1800-Present: http://www.sjsu.edu/faculty/harris/BritLitSurvey_F10/Engl56B_Frame.htm

For this course, we're not practicing any Digital Humanities, but we are looking at Digital Literature in the continuum of the survey, which is difficult considering we're still figuring out what that means. This is a lower-division English major's requirement, which means students typically haven't had the requisite course on how to evaluate literature in various genres.

All three of these courses offer students a sense of play. The most Digital Humanities-focused course, the Honors Digital Literature course that I'm teaching right now, even demands the play defined above, that unstructured, imaginative playing that happens with tag or hide n'go seek. In fact, when we started the section on e-literature, I provided no rules for close reading; my literature majors hated that and still are admonishing me for not supplying them with the tools. This is where English/literary studies have done them wrong (and me for that matter). We train our students

with a set of distinct rules, even asking some of them to master those rules. And, then and only then, do we allow them to break those rules. But, what if we never tell them the rules to begin with? What if we ask them to play first, explore, discover and then we provide them with a set of rules? I'm talking about reversing the curriculum to privilege bottom-up pedagogy.

Though much of this long paper is based in anecdotes about students enrolled in a large, public university, I believe it provides evidence of the potential for altering traditional pedagogy using the very thing that Digital Humanities encapsulates: play.

The Colonial Despatches of Vancouver Island and British Columbia: a Digital Edition of a Large-Scale Document Collection

Holmes, Martin
mholmes@uvic.ca
University of Victoria

Shortreed-Webb, Kim
ksw@uvic.ca
University of Victoria

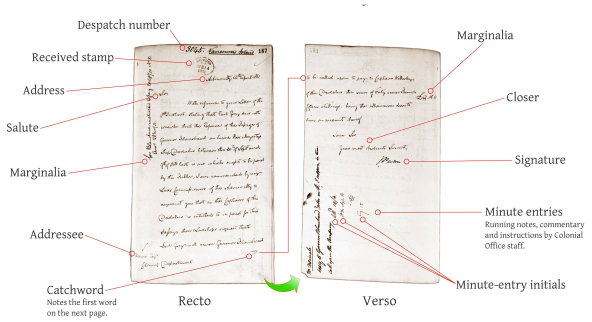
The modern Canadian province of British Columbia has been inhabited for at least 12,000 years, but its colonial history begins with the visits of Spanish explorers, in the 18th century, followed by voyages by Cook and Vancouver in the 1770s and 1790s. They were followed by other explorers, and by the Hudon's Bay Company, which established trading posts in the region, and became the *de facto* agents of British colonization until the formal establishment of the colony of Vancouver Island in 1849, and later the British Columbia colony in 1858. The corpus of historical texts in our collection *The Colonial Despatches of Vancouver Island and British Columbia* (`http://bcgenesis.uvic.ca/`) currently comprises over 7,000 documents, and covers the years between 1846, when negotiations began between the Colonial Office and the Hudson's Bay Company over the future of the territory, and 1871, when the young colony of British Columbia became a province in the Canadian federation.

Our documents have a somewhat troubled history. During the 1980s and 90s, a team led by James Hendrickson transcribed this huge collection into Waterloo Script, to produce a 28-volume print edition, which failed to find a publisher because of its scale. After Dr. Hendrickson retired, the markup files were largely forgotten, until they were rediscovered on an aging server which was scheduled to be shut down. At Digital Humanities 2008, we described how we retrieved the data and converted the Waterloo Script to TEI P5 (Holmes & Newton 2008).

Since then, the project has given birth to a full-scale digital edition. In addition to the original transcriptions, we now have over 18,000 page-images, with more

being added every week; these are being linked into the transcriptions at every page-break. We have also generated several hundred biographies of people mentioned in the despatches, along with short articles on nearly 200 places/locations and 100 ships which feature in the correspondence. In addition, we have acquired digital versions of more than 200 contemporary maps, many of which form part of the correspondence, and we will be adding many more in the coming months.

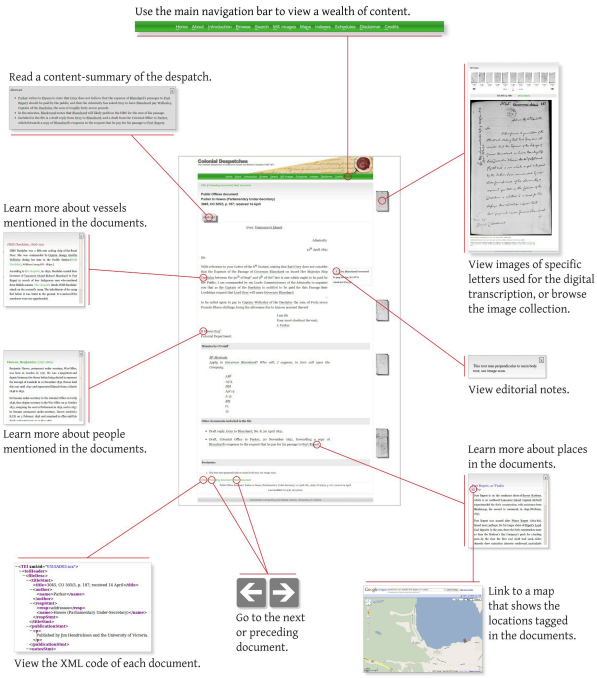
The documents themselves present significant difficulties for transcribers.



A simple despatch

This very short example shows some of the core features, and demonstrates the typical processes through which a despatch from Vancouver Island to London would go. Each despatch received in London would be logged in and assigned a number, then the Colonial Office staff would annotate it with a series of minutes recording their deliberations regarding the appropriate responses; often, letters would be written to other departments in the bureaucracy requesting guidance or information, and finally a reply would be drafted (Hendrickson 2008 describes this process in detail). At the end of the year, the despatches would be bound into one or more volumes for storage. The peripheral correspondence, including enclosures and attachments with the original despatch, are often bound up with the despatch itself, and the most significant have also been transcribed as part of our collection.

In our web application, we have attempted to reproduce all the pertinent features of the original text in a form which makes them more accessible to the reader, while providing easy access to our database of information on people, places, and ships:



A digitized despatch

The correspondence itself is rarely less than entertaining, and frequently exciting; it describes the difficulties encountered by a small, distant and relatively unsupported outpost of empire, struggling to deal with internal conflict, often-hostile First Nations groups, lawlessness, smuggling, and a somewhat threatening American population to the south. There are murders, shipwrecks, gold strikes and adventures of all kinds. However, the collection presents particular difficulties for readers, especially non-experts, arising out of the length of time required for a despatch to make its way between London and Victoria. Between the transmission of a despatch and the receipt of a response addressing its issues and questions, six months might elapse, and during that time, dozens or even hundreds more despatches would be sent. As a result, there is no apparent meaningful sequence in which to read the documents, as one might read an exchange of letters between two correspondents living closer to each other. A good search engine helps, of course, but we have also tried to provide other methods for users to navigate through the collection, through our markup of people, places and vessels. Each instance of one of these items in the text is linked to its "biography", and the biography can retrieve links to every mention of that person, place or ship anywhere in the correspondence, so it becomes possible to find paths through the documents based on these elements. We are also marking up some of the contemporary maps in the collection, using the *Image Markup Tool*, and integrating them with the

place database, so that it is possible to jump from the mention of a place inside a document to a "biography" of the place, and thence to the specific location of that place on any contemporary maps on which it appears, as well as Google Maps.

The entire collection (including the maps) is marked up in TEI P5 XML, and the web application is constructed using Apache Cocoon and the eXist XML database. Our poster presentation will deal with our approaches to markup, the web application architecture, and how we have attempted to overcome some of the challenges inherent in the scale and complexity of the collection.

References

Hendrickson, James E. (2008). 'The Colonial Office in 1858'. *The Despatches of British Columbia.* http://bcgenesis.uvic.ca/intro.htm#co_1858.

Holmes, M. and Newton, G. (2008). 'Rescuing old data: Case studies, tools and techniques'. *Digital Humanities 2008.* Pp. 127-131. http://www.ekl.oulu.fi/dh2008/Digital%20Humanities%202008%20Book%20of%20Abstracts.pdf.

NeDiMAH a Network for Digital Arts and Humanities

Hughes, Lorna
lorna.hughes@llgc.org.uk
National Library of Wales

Jannadis, Fotis
fotis.jannidis@uni-wuerzburg.de
Institute for German Philology, University of Würzburg

Schreibman, Susan
s.schreibman@ria.ie
Digital Humanities Observatory, Royal Irish Academy

---

1. Introduction

The European Science Foundation has recently recommended the NeDiMAH Network for funding through its 2009 Research Network Programme. The Network will run from June 2011-June 2015, and will provide a focus for researchers using digital research methods in the arts and humanities. NeDiMAH will provide a “methodological layer” to enhance and add value to infrastructure initiatives, and to digital collections. Outputs of the Network will provide evidence of the value of ICT methods for arts and humanities research. This poster will describe the Network and encourage wide participation amongst DH2011 attendees in Europe, and explore pan-national collaboration.

2. Aims of the Network

The NeDiMAH Network will examine the practice of, and evidence for, advanced ICT methods in the arts and humanities across Europe, and articulate these findings in a series of outputs and publications. NeDiMAH will provide a locus of networking and interdisciplinary exchange of expertise among the trans-European community of digital arts and humanities researchers, as well as those engaged with creating and curating scholarly and cultural heritage digital collections. NeDiMAH will work closely with the EC funded e-research infrastructure projects DARIAH (Digital Research Infrastructure for the Arts and Humanities, http://www.dariah.eu) and CLARIN (Common Language Resources and Technology Infrastructures, http://www.clarin.eu), as well as other national and international

initiatives. The Network will bring together practitioners to examine the use of formal computationally-based methods for the capture, investigation, analysis, study, modelling, presentation, dissemination, publication and evaluation of arts and humanities materials for research. This research will contribute to the classification and expression of ICT methods used in the arts and humanities in three key outputs: a map visualising the ICT methodological commons; an enhanced ICT Methods Ontology; and a collaborative forum for the European community of practitioners active in this area. These outputs will serve to formalize and codify the expression of work in the digital arts and humanities, give greater academic credibility to this work, and enable peer-reviewed scholarship in this area. NediMAH will maximise the value of national and international e-research infrastucture initiatives by developing a methodological layer that allows arts and humanities researchers to develop, refine and share research methods that allow them to create and make best use digital methods and collections. Better contextualization of ICT Methods will also build human capacity, and be of particular benefit for early stage researchers.

3. Background

Advanced ICT methods for discovering, annotating, comparing, referring, sampling, illustrating, and representing digital content (Unsworth, 2000) can be found at a key point of intersection between disciplines, collections and researchers: data-rich disciplines (e.g. archeology, library and information science, and musicology) have refined new ICT methods, and within the data-driven sciences research methods have emerged around data and information processes. The use of advanced ICT methods can effect significant benefits in arts and humanities scholarship. Humanities research can also benefit from the significant volume of digital material available to arts and humanities researchers, the access and use of which is now being supported by the development of research infrastructures.

Despite this activity, uptake and impact of ICT based methods remains fragmented. A recent DARIAH investigation into research practice and Research Actors, examining the institutional settings of digital scholarship in the arts and humanities, has shown that the use of ICT research methods is often concentrated in specific academic disciplines, or in libraries or archives, and there are few opportunities for transfer of knowledge across disciplinary boundaries. This creates disciplinary "silos", and communities of practice tend to develop around disciplines, rather than research methods (e.g., archeological computing, etc). Reasons for limited uptake and multidisciplinary collaboration were expressed in the ACO*HUM report (`http://gandalf.uib.no/AcoHum/`), and in the recent evaluation of the UK's AHRC ICT Methods Network (`http://www.methodsnetwork.ac.uk/evaluation`). These indicated that there is an urgent requirement for an international collaborative effort to undertake a formal analysis and expression of the ICT methods that can be used for arts and humanities research. Computational methods demand the utmost rigour and precision in their application, and accordingly, research practitioners working in the emerging field of the digital humanities have begun to formalize new theories of the interaction between content, analytical and interpretative tools and technologies, methodological approaches, and disciplinary kinships.

NediMAH will be an interdisciplinary, international Network of expert practitioners in the digital arts and humanities with the following objectives:

- To investigate and articulate the use of formal computationally-based methods for the capture, investigation, analysis, study, modelling, presentation, dissemination, publication and evaluation of arts and humanities materials for research.
- To facilitate collaboration in this research by:
  i. Building a community of practice that is inclusive in terms of disciplinary coverage and national representation, as well as seeking the active participation of scholars at all stages of the career cycle
  ii. Developing a framework for common exchange of expertise and knowledge
  iii. Linking researchers with their peers across the disciplines
  iv. Enabling participants to develop, share and refine ICT methods as the core elements of digital scholarship and articulate these methods formally
- To map the outputs and findings of these investigations in two digital resources:
- A "methodological commons" to express disciplinary commons, partnerships and synergies, and the potential for cross and interdisciplinary partnerships.
- The ICT Methods Taxonomy:
  i. To build a community knowledge base by extending the arts-humanities.net and

DRAPIER resources across Europe. This resource will be embedded into DARIAH beyond the end of the funding period to ensure sustainability and continued dissemination of Network outputs.

ii. To investigate issues related to the scholarly publishing of ICT methods in the arts and humanities

iii. To publish Network research in a series of books and articles.

4. Network Activities

NeDiMAH has a collaborative structure that is trans-European, interdisciplinary, and able to leverage existing nationally funded research and research support activities that can collectively support the development of a better understanding of the role of advanced ICT methods in arts and humanities research. This will support a series of core activities and a mechanism for exchange of expertise and material. These activities include a series of Methodological Working Groups, convened to consider specific methodological areas over the entire duration of the Network. They will investigate the topic from three areas of scientific focus:

1. Investigating the use of the method and gathering information about specific projects that use it
2. Analysis of current practice
3. Modelling ways in which the method can be applied across the disciplines in scholarly practice

Each Working Group will convene a Workshop each year, a key activity by which advanced methods will be formalized.

Proposed Working Group topics include:

1. Spatial and Temporal Modelling
   ICT methods include agent-based modelling geo-temporal referencing and predictive spatial modeling.
2. Information Visualization
   Visualisation refers to techniques used to summarise and present data visually, in a form that enables people to understand and analyse the information. Formats can include images (3-D or 2-D), maps, timelines, graphs and tables.
3. Ontological methods
   Ontological mapping is used to semantically interrelate information from diverse sources to represent complex relationships. In order to do that, it relies on ontologies, formal representations of a set of concepts and relations.
4. Information extraction and data mining
   Text and data mining can reveal new knowledge from (usually) larger amounts of textual data extracting hidden patterns, analysing the results and summarising them into a useful format.
5. Linguistic Corpora for interdisciplinary research
   To study language as expressed in such corpora, corpus linguistics makes use of methods such as annotation, content analysis and parsing. Originally done by hand, corpora are now largely derived by an automated process.
6. Methods and tools for working with digital manuscripts and images
   Methods include text encoding to image restoration, and tools for digital editions

5. Participation in the Network

NeDiMAH will invite the broadest participation from European researchers in the digital arts and humanities through an open call. The Network also has a Global Dimension, which would enable the participation of researchers outside the EU. NeDiMAH has the full support of the following international digital humanities organizations: the Allied Digital Humanities Organizations (ADHO), the Association of Literary and Linguistic Computing (ALLC); and the Association of Computing in the Humanities (ACH) and CentreNet. Digital Humanities 2011 will be an important opportunity for the Network to invite collaboration.

Visualization of Co-occurrence Relationships Using the Historical Persons and Locational Names from Historical Documents

Itsubo, Sho
cm001061@ed.ritsumei.ac.jp
Graduate School of Science and Engineering, Ritsumeikan University, Japan

Osaki, Takahiko
cm003060@ed.ritsumei.ac.jp
Graduate School of Science and Engineering, Ritsumeikan University, Japan

Kimura, Fuminori
fkimura@is.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

Tezuka, Taro
tezuka@media.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

Maeda, Akira
amaeda@media.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

1. Introduction

In recent years, there is an increasing use of digital technology in the study of humanities. Many historical documents are now digitally archived, enabling further analysis using computers. There are archives that are accessible on the World Wide Web, including Union Catalog of the Collections of the National Art Museums, Japan (Independent Administrative Institution National Museum of Art, 2010) and Perseus Digital Library (Crane, 2011).

Until recently, the storage of historical documents and data has been the main target of digital archive research. There are, however, many works that go on to analyze the content of the historical documents using text mining techniques. In this paper, we propose a method of visualizing relationships among historical persons using personal names and place names appearing in documents.

2. Proposed Methods

We propose two methods to extract and visualize the relationships among persons from historical documents. The goal of two methods is to extract dynamics of relationships among historical persons. The first method tracks temporal change in a relationship and visualizes it. The second method utilizes locational information to obtain latent relationships among persons based on their spatial activities.

3. Personal Relationships Using the Co-occurrence Information between Persons

We use three historical documents in Japan, “Azumakagami”, “Gyokuyou” and “Hyohanki”. “Azumakagami” is an official record written in Kamakura period (A.D.1180-1266). “Gyokuyou” is a personal diary written by Kanezane Kujou (A.D.1164-1200) in the late Heian period from early Kamakura period. “Hyohanki” is a personal diary written by Nobunori Taira (A.D.1112-1187) in late Heian period. We first extracted persons’ names using “Azumakagami” and “Gyokuyou” databases (Fukuda, 2002) that contains an index of persons. This is because in historical Japanese documents, persons are often referred to using other names besides their real names. The index of persons includes such cases also. Since the three historical documents are diaries, we defined co-occurrence to be the case where two persons appearing in the same date. We calculated co-occurrence frequencies among persons for each year.

4. Personal Relationships Using the Co-occurrence Information between Person and Locational Information

In this method, we used “Hyohanki”. In the first step, we obtain frequencies of co-occurrences between each person’s name and place names. Since “Hyohanki” is written in ancient Japanese, we cannot attach part-of-speech tags using existing morphological analyzers that were trained using modern Japanese. We therefore used pattern matching to find place names that were included in the dictionary. We used the “Index of Kyoto’s Place Names” created by Noboru Tani based on the “Outline of Heiankyo” (Tsunoda, 1994) and “Japan’s Historical Place Names 27: Place Names of Kyoto” (Hayashiya et al., 1979) as the data sources for place names.

We use co-occurrence as the indicator of the relationship between a person and a location. If a

person's name and a place name appear in the same paragraph, we consider it as a co-occurrence between the two. Since each paragraph often covers a specific situation or a topic, we consider it to be a better unit than dates.

Each place name is considered as a dimension of a vector space. For each person, we create a vector having the number of co-occurrences with a place name as the component. For the similarity measure, we used cosine similarity.

We use the similarity measure and the result of clustering for visualization. We used JUNG, a Java open source library for drawing graph structure.

5. Results of Visualization and Discussion

The result of visualizing temporal change in a relationship

Figure 1-3 show the results of visualization using the method described in Subsection 2.1. Figure 1 shows the change in the relationships of Yoritomo Minamoto and Yoshitsune Minamoto to Emperor Goshirakawa for seven years (from 1184 to 1190).

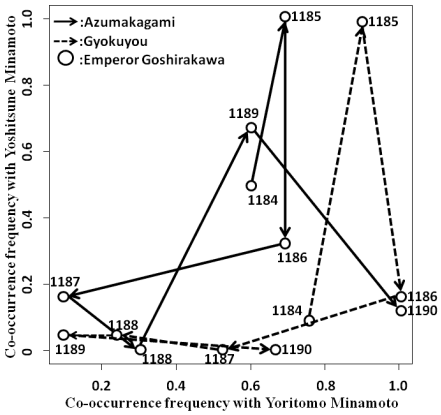


Fig. 1. Transition of relationships of Yoritomo Minamoto and Yoshitsune Minamoto to Emperor Goshirakawa

The horizontal axis is the co-occurrence frequency between Emperor Goshirakawa and Yoritomo Minamoto, who became Shogun (the leader of samurai warriors) in 1192. The vertical axis is the co-occurrence frequency between Emperor Goshirakawa and Yoshitsune Minamoto, who is a younger brother and the archrival of Yoritomo. The arrows indicate the transition based on "Azumakagami", while the dot arrows indicate the transition based on "Gyokuyou".

The result illustrates the change in the relationship among Yoritomo, Yoshitsune and Emperor Goshirakawa. The change in 1184-1187 indicated by the arrow closely resembles that of the blue allow. The change also matches well with the historical fact. Transitions in 1187-1190 are different among two documents. This is assumed to be due to the fact that "Azumakagami" is a diary written by the Shogunate side and "Gyokuyou" is a diary written by the Imperial court side. The relationship between Yoritomo Minamoto and Yoshitsune Minamoto, who were both samurai warriors, are probably not described well enough in "Gyokuyou".

Figure 2 shows the change in the relationships of Yoritomo Minamoto and Emperor Goshirakawa to Yoshiyasu Ichijou for seven years (from 1185 to 1191).

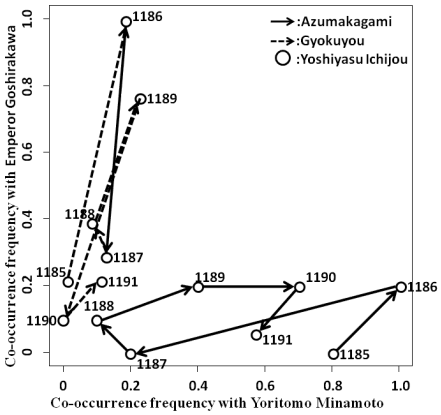


Fig. 2. Transition of relationships of Yoritomo Minamoto and Emperor Goshirakawa to Yoshiyasu Ichijou

The horizontal axis is the co-occurrence frequency between Yoshiyasu Ichijou and Yoritomo Minamoto. The vertical axis is the co-occurrence frequency between Yoshiyasu Ichijou and Emperor Goshirakawa. The arrows indicate the transition based on "Azumakagami", while the dot arrows indicate the transition based on "Gyokuyou".

The result illustrates the change in the relationship among Yoritomo, Emperor Goshirakawa and Yoshiyasu. Yoshiyasu is a person who was active on the emperor side and the samurai side. Therefore, the relation between Yoritomo and Emperor Goshirakawa is strong. However, the change is different in "Azumakagami" and "Gyokuyou"."Azumakagami" is the content of the samurai side and "Gyokuyou" is the content of the emperor side. Figure 2 shows the feature of two historical documents clearly.

Figure 3 shows the change in the relationships of Kiyomori Taira and Emperor Goshirakawa to Motofusa Fujiwara for the six years (from 1166 to 1171).

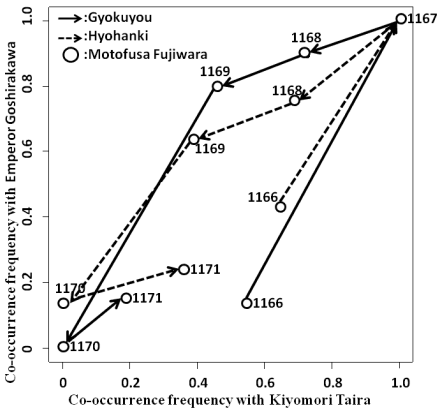


Fig. 3. Transition of relationships of Kiyomori Taira and Emperor Goshirakawa to Motofusa Fujiwara

The horizontal axis is the co-occurrence frequency between Motofusa Fujiwara and Kiyomori Taira. The vertical axis is the co-occurrence frequency between Motofusa Fujiwara and Emperor Goshirakawa. The arrows indicate the transition based on “Gyokuyou”, while the dot arrows indicate the transition based on “Hyohanki”.

The result illustrates the change in the relationship among Kiyomori, Emperor Goshirakawa and Motofusa. “Gyokuyou” is the content of the emperor side and the author of “Hyohanki” is a person on the emperor side. Therefore, the contents of two historical documents are similar and the transition of Figure 3 is also similar. The results suggest the possibility of estimating author's standpoint from the document.

6. The Result of Visualizing Latent Relationships using Locational Information

Using our proposed method of 2.2, we created graphs that visualize relationships between historical persons. We focused on the time range of Hougen Rebellion, starting in early July 1156 and ending in late July of the same year.

Hougen Rebellion is a short civil war caused by a power struggle between Emperor Goshirakawa and former Emperor Sutoku.

We chose 78 persons belonging to either the faction following former Emperor Sutoku or the faction following Emperor Goshirakawa (Hyohanki Reading Circle, 2007). Most of them are aristocrats and samurai warriors. It is distinguishable from historical records to which faction each person belonged to. In “Hyohanki”, 31 of these persons had co-occurrence with location names. We used K = 3 for K-means clustering and L = 20 for initialization. K is the number of clusters used in K-means clustering. L is the number of repetitions for finding the optimal initial centroids.

Figure 4 shows the result of visualization using the similarity of co-occurring location names.

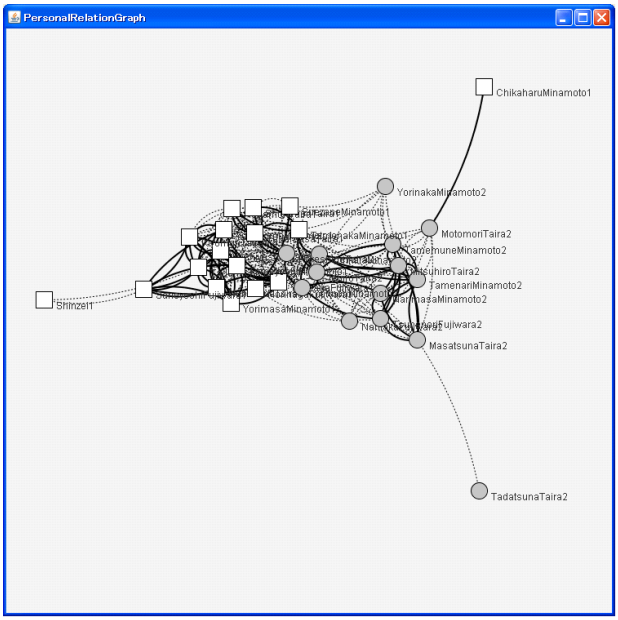


Fig. 4. Relationships among persons and historical factions

The number next to the node and the color indicates to which faction each person belonged to. A node labeled “1” (box) indicates that he followed Sutoku. On the other hand, a node labeled “2” (circle) indicates that he followed Goshirakawa. Lines are drawn when similarity is over 0.4. Dotted lines indicate similarity between 0.4 and 0.7. Solid lines indicate similarity over 0.7.

Figure 5 shows the result of clustering. The number next to each node indicates to which cluster the person was allocated to.

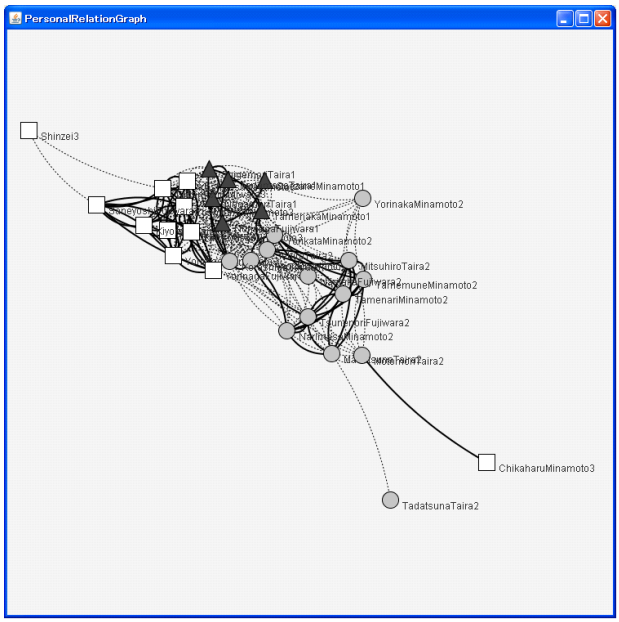


Fig. 5. Result of clustering persons using locational information

Colors also indicate clusters. The cluster 1 is box, the cluster 2 is circle, and the cluster 3 is triangle. Table 1 shows to which cluster and to which faction each person belonged to.

	Cluster 1	Cluster 2	Cluster 3
faction of former Emperor Sutoku	Nagamori Taira Norinaga Fujiwara Tamenaka Minamoto Tadamasa Taira	Narimasa Minamoto, Tadatsuna Taira Yorikata Minamoto, Yorinaka Minamoto Tamenari Minamoto, Yorinori Minamoto Tsunenori Fujiwara, Tadatsuna Taira Tamemune Minamoto, Iehiro TairaMitsuhiro Taira, Naritaka Fujiwara	Yorinaga Fujiwara Chikaharu Minamoto
faction of Emperor Goshirakawa	Suezane Minamoto Shigemori Taira Koreshige Taira	Motomori Taira	Shinzei Yoshitomo Minamoto Yorimasa Minamoto Tameyoshi Minamoto Saneyoshi Fujiwara Kiyomori Taira Yoshiyasu Minamoto Tadamichi Fujiwara Nobukane Taira

Table. 1. Comparisons of factions and clusters

We proposed a method of revealing and visualizing relationships among historical persons by focusing on place names appearing in digitized historical documents. We used cosine similarity and a modified K-means algorithm to create graphs and cluster persons.

In the experiments, we used persons that we know to which faction he belonged to during the Hougen Rebellion. The result showed a strong correspondence between the factions and the clusters, indicating effectiveness of using location information for clustering people.

7. Conclusion

The results of our experiments showed that relationships between historical persons can be extracted using historical documents. The experiment described in Subsection 3.1 showed that temporal change of a relationship can be visualized using change in co-occurrence frequencies. The experiment described in Subsection 3.2 indicated a strong correspondence between the factions and the clusters, indicating effectiveness of using location information for clustering people.

References

Independent Administrative Institution National Museum of Art (2010). *Union Catalog of the Collections of the National Art Museums.* `http://search.artmuseums.go.jp/`.

Crane,G.R. (2011). *Perseus Digital Library.* `http://www.perseus.tufts.edu/`.

Fukuda,T. (ed.) (2002). *Azumakagami-Gyokuyou Database*

Tsunoda,B. (ed.) (1994). *The Outline of Heiankyo*

Hayashiya,T., Murai,Y., Moriya,K. (1979). *Japan's Historical Place Names 27: Place Names of Kyoto*

Sakai,M., Yamada, S., Onoda,T. (2010). *Initialization of k-means method using independent component analysis. The 24th annual meeting of the Japanese Society for Artificial Intelligence*

Hyohanki Reading Circle (2007). *The Index of Hyohanki's Persons' Names*

The Effect of Cheating on Player Engagement in Video Games

Keenan, Andy
andrewtkeenan@gmail.com
Humanities Computing, University of Alberta

1. Introduction

Creating an engaging video game requires the appropriate balance between challenge and reward for players. This balance is known as “difficulty”. Currently, difficulty is created, designed, and managed by video game developers. Difficulty is controlled by fundamental design decisions. For example, in Super Mario Brothers (NES 1986), Mario has a specific number of lives. When Mario loses those lives by making mistakes, the game is over. The difficulty of Super Mario Brothers is progressing through the levels without losing lives. The game balances the challenge of limited lives with the reward of completion. If Super Mario Brothers failed to find the right difficulty, the game would be either too frustrating or too easy. Finding the appropriate level of difficulty for the largest number of players is an important design decision made by video game developers.

The purpose of this project is to explore player-controlled approaches to difficulty in video games. This project repositions the existing power relationship between player and developer, where the developer decides the difficulty. Based on my previous research, I will be analyzing how cheating allows players to manipulate difficulty and change the relationship between player and developer. Cheating enables players to alter game difficulty, enabling players to find their own balance between challenge and reward. This project will provide design recommendations for the video game industry to re-imagine the relationship between video game player and video game developer, empowering the player to determine the rules of their game play experience.

2. Context

Johan Huizinga argues that play creates a space apart from normal life. He referred to this space as the magic circle. While in play, players are subject to a special set of rules. Play is dependent on these rules. Participating in play requires all players to enter into the magic circle and abide by its rules. The stability of the game depends on its rules: “[if] the rules are transgressed the whole play-world collapses. The game is over” (Huizinga 1950, p3). Huizinga’s magic circle requires a second examination in regards to video games. Allowing players to manipulate the rules alters the power relationship between developers and players, which could enable players to create more engaging experiences.

Several academic studies have attempted to determine what makes games engaging. Cognitive research in video games suggests that game developers must adjust cognitive difficulty requiring players to change their cognitive model to succeed. This will keep players engaged by challenging their cognitive process (Graham, Zheng & Gonzalez 2006). Other studies found that players require realistic worlds, intuitive controls, character customization, exploration, and unpredictability (Wood, Griffiths, Chappell & Davies 2004). A player-feedback study suggests that a combination of different intensities of challenge, combining “hard fun” (complex strategy, difficult challenges, and powerful enemies) and “easy fun” (exploration, simple puzzles, and novel experience) creates an ideal experience (Lazarro 2004). Yet another study argues for the importance of goal-oriented play with few negative consequences (Provenzo 1991). More abstractly, play must be internally motivated, simultaneously transcend and reflect reality, focus on the process over the result, and provide safe yet unpredictable experience (Stagnitti 2004). There is a gap in the current literature exploring the impact of cheating on engagement.

3. Thesis

This presentation explores the relationship between cheating and engagement in video games. Recent innovations in video game design allow players to manipulate game difficulty through time manipulation, an activity once considered cheating. These time manipulation games serve as a case-study to examine cheating and the effect on player engagement. Manipulating time allows players to control difficulty and find an iterative balance between challenge and reward. Players can also create emergent game play types by deciding what constitutes a meaningful “beat” of play.

4. Methodology

Using a mixed methods approach, I combined a heuristic inquiry of games that allow time manipulation with Foucauldian discourse analysis. I examined the practice of time manipulation in several console

video games including Forza Motorsport 3 (Microsoft 2009); Braid (Microsoft 2008); Skate (Electronic Arts 2007), Skate 2 (Electronic Arts 2009); World of Goo (2D Boy 2008); Prince of Persia: Sands of Time (Ubisoft 2003); Demon's Souls (Sony 2009); and Madden 09 (Electronic Arts 2008). Heuristic inquiry is a qualitative research approach concerned directly with human knowing and self-inquiry. This method "is aimed at discovering the nature and meaning of an experience" (Hiles 2008, p389). This is a departure from mainstream research "in that it explicitly acknowledges the involvement of the researcher to the extent that the lived experience of the researcher becomes the main focus of the research" (ibid). Based on my experiences with these games, I conducted a Foucauldian discourse analysis focusing on player engagement and player empowerment. I analyzed how activities once considered cheating altered the player's relationship to the game.

5. Conclusion

I discovered that player-determined difficulty effected by level of engagement with the games. By controlling the game's difficulty through its interface, the level of engagement was increased. Manipulating time allows players to learn iteratively and manage their level of challenge in the game experience. As an interface feature, manipulating time encouraged the "flow" state: Mihaly Csikszentmihalyi's theory of optimal experience. Csikszentmihalyi's *Flow: The Psychology of Optimal Experience* (1990) describes flow as being completely involved in an activity. Reducing frustration and allowing players to control their experience by rewinding time creates this flow state in video games.

References

(2008). *Braid.* Number None: Microsoft.

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience.* New York: Harper and Row.

(2009). *Demon's Souls.* Atlus: Sony.

(2009). *Forza Motorsport 3.* Turn 9: Microsoft.

Foucault, Michel (1972). *The Archaeology of Knowledge and the Discourse on Language. (A. Sheridan, Trans.).* New York: Pantheon.

Graham, J., Zheng, L., and Gonzalez, C. (2006). 'A Cognitive Approach to Game Usability and Design: Mental Model Development in Novice Real-Time Strategy Gamers'. *CyberPsychology & Behaviour.* Volume 9, Issue 3: 361-366.

Hiles, David (2008). 'Heuristic Inquiry'. *The Sage Encyclopedia of Qualitative Research Methods.* Lisa Given (ed.). Los Angeles: Sage, pp. 389-392.

Huizinga, Johan (1950). *Homo ludens: A study of the play element in culture.* Boston: Beacon Press.

Lazzaro, N. (2004). *Why We Play Games: Four Keys to More Emotion Without Story. Player Experience Research and Design for Mass Market Interactive Entertainment.* Oakland, CA: XEODesign Inc...

(2008). *Madden 09.* EA Tiburon: Electronic Arts.

(2003). *Prince of Persia: Sands of Time.* Ubisoft Montreal: Ubisoft.

Provenzo, E. F. (1991). *Video kids: Making sense of Nintendo.* Cambridge, MA: Harvard University Press.

(2007). *Skate.* EA Black Box: Electronic Arts.

(1986). *Super Mario Brothers.* Nintendo: Nintendo Entertainment.

Stagnitti, K. (2004). 'Understanding play: The implications for play assessment'. *Australian Occupational Therapy Journal.* Volume 51, Issue 1: 3-12.

Wood, R., Griffiths, M., Chappell, D., and Davies, M. (2004). 'The Structural Characteristics of Video Games: A Psycho-Structural Analysis'. *CyberPsychology & Behaviour.* Volume 7, Issue 1: 1-10.

(2008). *World of Goo.* 2D Boy: 2D Boy.

Between Close and Distant: Historical Editing Methods at Intermediate Scale

Knox, Douglas W.
knoxdw@gmail.com
Newberry Library

Between mass digitization of millions of images and texts and intensive scholarly work on close digital representation, in the humanities there are many collections of texts at intermediate scales that require their own strategies for management, editing, and analysis. The Text Encoding Initiative grew out of scholarly needs in working with digital texts that were encoded manually with considerable editorial care. In recent years it has become possible to begin to address the challenge of working with digital resources at scales that had been difficult to imagine previously —a challenge captured in Gregory Crane's memorable question, "What Do You Do with a Million Books?" [1]

More recently Crane seemed to reply to his own question with the demand, "Give us editors!," recognizing the need for new kinds of editing and scholarship that will combine the strengths of digital methods and domain-appropriate human judgment [2]. Between markup of single texts and mass digitization of millions of books, increasing attention is called for at intermediate scale. The present project is a case study of one modest example of this sort of work.

At the scale of tens of thousands of texts, manual editing methods are impractical or not cost-effective, yet there is often still considerable value in structured markup to support management, discovery, and scholarly use beyond what uncorrected OCR or plain text can support. While this is not in itself a novel observation, thinking about interpretive issues and scale in relation to digital editing has in general been much more developed with respect to literary texts than it has been for nonliterary historical sources.

The poster will outline some of editing questions and processes that can arise in applying digital methods to a large aggregation of related historical documents, drawing on examples from the Chicago Foreign Language Press Survey, an NEH-funded project of the Newberry Library that is creating an electronic publication and database using relatively simple TEI structures to represent a collection of thousands of newspaper articles prepared in the 1930s. The original Press Survey was a project of the U.S. Works Progress Administration in the 1930s that selected, translated into English, edited, and organized articles published in Chicago from 1861 to 1938 in newspapers in twenty-two linguistic and ethnic groups. The project produced 120,000 half-sheets of typescript, recording what we now know are more than 48,000 articles.

The present-day Press Survey project worked with a vendor to produce a set of simple TEI transcription files from images digitized from microfilm by the library at the University of Illinois at Urbana-Champaign. The body of each article has been transcribed into paragraphs and occasionally tables, without further markup below that level. Considerable effort, however, went into making sure the files accurately capture the structure of information that constitute metadata for each article: ethnic group, primary and secondary subject code classifications, source, date, and title. At the publication end of the process, web developers populate a database with metadata drawn from edited TEI files and create a searchable, browseable interface for the collection.

Between the vendor's XML and the web presentation is an essential set of editing tasks. Many editing steps, beginning with XML validation, apply to single documents, and scale up through simple repetition, preferably automated repetition with manual exception handling as required. Often this is a matter of imagining what might have gone wrong, and then devising tests for it. Are any items, pages, or essential data structures missing? Are there paragraphs that are suspiciously short? Are there any duplicate values one would not expect, whether identifiers, file name references, or page numbers? Is there data integrity in expected relations between file names and internal identifiers? Is there the sequential continuity we would expect in any numbering or other patterns where the sequence of the original material implies some kind of consistency or order in metadata fields?

Some of the most interesting editing tasks, however, are those that relate to the collection as a whole, and require an iterative exploratory process. The 1930s editors of the Press Survey thought carefully about their choices and methods in organizing a large body of material. They invented their own hierarchical subject code scheme, and the linear arrangement of the typescript indicates how they prioritized ethnic groups and primary subject codes over date of publication and secondary subject terms. They faced inevitable limitations in trying to managing a textual database of tens of thousands of records with nothing more than paper, procedural guidelines, and human

attention. For seventy years, in the absence of a digital representation of the editorial model of the WPA editors, it was impossible to get a full picture of the choices they made. Now, with digital methods, we can better enjoy the benefit of their work and yet also see the limitations of their methods, even according to their own intentions, better than they themselves were able to.

While scholars have made good use of typescript, microfilm, and digital images, until the Press Survey was modeled as a database of articles its aggregate characteristics were obscure. No index recorded the range and distribution of values in important metadata fields, including dates and the names of newspapers and other sources. Creating a digital representation now requires simultaneous analysis and editorial decision-making. How often did the original editors depart from their own advertised controlled vocabularies? What simple errors should now be corrected, what normalization of data would best serve the current digital project? Do the data structures chosen in anticipation of full-scale digitization fit what we can now know about the documents in the aggregate? With an eye to both central tendencies and odd but perhaps telling exceptions, how can editing choices mediate a digital collection for efficient use while also preserving some necessary kinds of transparent resistance to user expectations, features that may be essential to its character as a complex historical primary source?

The poster will be organized around these kinds of questions and tasks, with a series of illustrative case studies drawn from Foreign Language Press Survey documents and data.

Generalizing from this case, the primary point is that intermediate scale matters in its own right. If a collection is interesting as such, it will likely require more than the repeated application of the kind of editing applied to individual texts, and it may also require decision-making beyond the methods and standards designed to provide access to aggregations at much larger scale. We must be able to see both the forest and the trees.

References

Crane, G. (2010). 'Give us editors! Re-inventing the edition and re-thinking the humanities'. *Connexions.* http://cnx.org/content/m34316/ (accessed March 9, 2011).

Crane, G. (2006). 'What Do You Do with a Million Books?'. *D-Lib Magazine,.* V. 12. http://www.dlib.org/dlib/march06/crane/03crane.html (accessed March 9, 2011).

TEI Consortium (ed.) (2007). *TEI P5. Guidelines for Electronic Text Encoding and Interchange.* TEI Consortium. http://www.tei-c.org/Guidelines/P5 (accessed March 9, 2011).

Notes

1. (Crane, 2006).
2. (Crane 2010)

Roots of Performatology: From Uber-Marionette to Embodied Performative Agent

Maraffi, Christopher
topherm@soe.ucsc.edu
University of California, Santa Cruz

This poster presents humanities research the author completed in Digital Arts and New Media that has led to the development of a novel high-level Performatology approach to designing embodied agents in Computer Science. Agent research for interactive narrative and games have incorporated some performative dramatic theory (Mateas, 2004; Seif El-Nasr, 2004; Tannenbaum, 2008; Perlin, 1996), but being primarily influenced by literary approaches (Laurel, 1991; Murray, 1998; Austin, 1962; Searle, 1969), the focus has been towards developing a Neo-Aristotelian Poetics approach to interactive drama. Some agent work has been done on modeling improvisational performers from Theatre Arts (Magerko, 2010), but their micro-agent designs did not incorporate embodied gesture. Although additional research has been done on modeling the gesture of professional speakers for enhancing the personality of embodied conversational agents (Neff, 2008), little work has been done to procedurally model professional performers from the arts to enhance the gestural quality of embodied agents.

By and large, the emphasis for previous interactive narrative research has been to provide story authoring tools rather than tools for embodied performers to represent their craft in computational media. Avatars, Non-Player Characters (NPCs), and Intelligent Virtual Agents (IVAs), all tend to be designed to function primarily as embodied conversational agents, with gestural performance being supportive to speech acts that drive the narrative forward. Thus, interactive drama in games is in stark contrast to how drama developed in both classical live theatre and moving pictures, where gestural performance historically preceded speech acts, and where physical drama and visual spectacle dominated textual narrative.

When cinema started, the camera was the entry point for professional performers to migrate out of the proscenium and into the screen mediums of film and animation. Actors effectively digitized their embodied 3D gesture into animated 2D representations that translated their already developed fictive personas into the plastic time and space of analog media, where the characters could repeat the original performances indefinitely, even after the death of the performers. Multiple takes, editing, and visual effects allowed them to iteratively improve their linear performances for audiences. The dramatic personas or icons created by media stars of the last century arguably eclipse the narrative elements in any single story, and indicate a development path for embodied performative agents in New Media today.

2. From Uber-Marionette to Avatar Theatre: Imagining the Ideal Performer

Interactive drama has the potential to extend the actor's craft further than all previous acting mediums, but due to a technological divide, actors have been prevented from playing on the new stage. The current absence of performing artists in computational media, as well as the potential solution to the problem in a procedural performer, was anticipated by theater practitioner Edward Gordon Craig at the turn of the last century. An influential British theorist working out of Florence, Craig self-published his controversial theatre reformation opinions in a periodical called The Mask (1907-1929). He advocated the rise of the Director-Designer as a visionary artist in charge of all aspects of production, over both the author and players, and in this role he faced character believability problems that caused him to question the viability of the actor as an artistic medium. Craig particularly struggled with the unpredictable personalities of live actors in his productions, which he attributed to a fundamental problem of belief in the theatre. He conceptualized the perfect actor as one who had a single-minded belief in the idea of the character, in the ideal Platonic sense, which would in turn make the audience believe in the characterization enough to have an emotional response (Craig, 1963).

A former actor himself, Craig's views on acting as a belief problem was influenced by Neo-classical ritual acting techniques from masked theatre and puppetry. These art forms also moved him towards symbolic gesture, along with contemporary influences such as the stylized movement of actor Henry Irving and dancer Isadora Duncan, as well as the anti-realism of the Symbolists (Ayat-Confino, 1987). His innovative productions were larger than life spectacles of moving lights and set pieces, which he intended as a new type of immersive Kinetic Stage. Since he was working against realism or naturalism, but still wanted to create a believable experience for the spectator, designing a consistent theatrical experience was critical to him.

But the one thing Craig couldn't control was the live actors in his productions, who he claimed did not have the disciplined belief required to reliably portray his desired characterizations. So in 1907 he published his infamous essay *The Actor and The Uber-Marionette*, where he proposed a technological solution to his acting problem, and proclaimed that for the artistic future of the theatre, all live actors should be replaced by autonomous puppets (Craig, 1907).

In his essay, Craig compared acting to other art forms, and found it came up short. His main complaint was against the live actor as a performing medium. When playing in front of an audience, Craig claimed that actors allowed their personality or mind to get in the way of their characterization. Through either nervousness or ego, they would tend to frequently act out of character, breaking the believability of their portrayal for the audience. Painters and musicians, as visionary artists, could abstract and refine their art forms with complete control. Actors, who often relied on their own personality to carry a role, in Craig's opinion, did not qualify as artists. His ingenious solution was to remove the actor entirely from playing in front of a live audience. He proposed, for the good of the theatre, that actors had to be banned from the stage and replaced with an autonomous puppet he called the Uber-Marionette. Craig already used techniques from masked theatre and puppetry to insert a performing object between the actor and audience, which was known to create a distancing effect for the performer. The Uber-Marionette concept was an extension of these techniques intended to entirely remove the actor in form and personality, leaving only a refined version of their gesture as an embodied performative representation.

Craig's argument may be sound, but his purposely provocative delivery didn't receive a positive response from the theatre community, especially from actors. So there is some theater, to champion the inclusion of live performers in the field of computational media today. Clearly ahead of his time, and highly perfomative in writing The Mask (Taxidou, 1998), he was often misunderstood by his contemporaries who either took his words at face value, or assumed he was writing in metaphors.The argument can be made that Craig was actually proposing a vision of the ideal actor intended for a future performance medium not yet invented. He said as much in his Uber-Marionette essay when he wrote, "If you can find in Nature a new material, one which has never yet been used by man to give form to his thoughts, then you can say that you are on the high road towards creating a new art. For you have found that by which you can create it. It only remains for you to begin. The Theatre, as I see it, has yet to find that material." (Craig, 1907). Though technologically impossible in his lifetime, Craig never gave up trying to realize his concept, nor did he ever admit that it was impossible to build. Though highly criticized by his contemporaries, he showed an unshakeable belief that it would be invented someday by discovering a new control mechanism, "What the wires of the Uber-Marionette shall be, what shall guide him, who can say?" (Craig, 1963).

Although Craig's banishment of the actor has inadvertently been realized in computational media, the author contends that his vision of the perfect actor as an Uber-Marionette can also be realized by modeling the behavior of embodied agents on the gesture of live professional performers. If the essence of good acting is semiotic gestural technique, or a vector of poses and movements that convey a symbolic attitude of the intended persona, then the ideal acting medium is one that can iteratively refine a trained performer's gesture to a singular clear purpose in real-time for a live audience. Craig's imagined "wires" are procedural character algorithms trained on motion capture data, and his desired new "material" for Theatre is the virtual stage of games and interactive drama. The author's MFA thesis and accompanying Avatar Theatre performances were intended to translate Craig's vision to a New Media context, and to informally assess contemporary audience reaction to seeing a live performer interacting with a 3D character in a shared performance space (See Figure 1; Maraffi, 2010).

Figure 1

Audience feedback suggested that, in the context of a live dramatic performance, believable interaction and expressive appeal can possibly solve Computer Science problems for embodied agents, such as passing a gestural Turing Test and offsetting the Uncanny Valley Effect. For instance, there were

performances when audience members believed that a remote performer was controlling the Avatar character, as well as a positive general audience response to comedic interaction that countered any strangeness from seeing the avatar mimicking the performer's natural movement.

3. From Avatar Theatre to Performatology: Getting the Performer In the Game

So how does the above MFA research apply to games and interactive drama? It is the proof-of-concept for developing a Performatology approach to designing embodied performative agents at UCSC's Computational Cinematics Studio. Using this approach the author has designed a novel performer modeling agent architecture, IMP*R*SONA, which uses machine learning and motion capture to build a performer profile from a knowledge base of procedural gesture prototypes (see Figure 2; Maraffi, 2011).

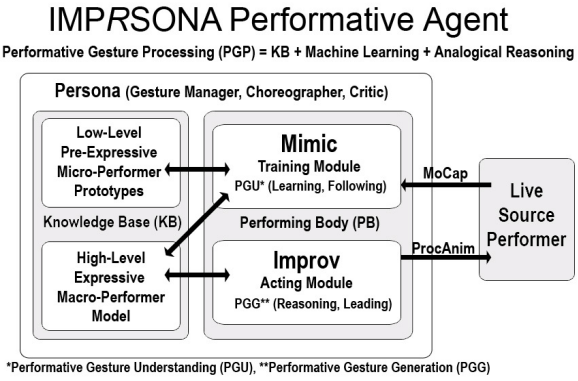


Figure 2

We are mapping a performative ontology from Performance Theory (Barba, 2005; Schechner, 2002; Aston, 1991), and from the author's experience as a performance artist and technical animator, to simulate the apprentice training process used by professional performers when learning and practicing their craft through mimicry and improvisation. The author's hypothesis is that the performer's technique, developed over many years of disciplined training, results in a semiotic quality of gesture that is integral to portraying believable, expressive, and appealing fictive characters.

Puppetry and animation over the last century have shown that abstracting principles from live performers can simulate the illusion of life in moving images (Thomas, 1981), as seen in Disney's Mickey, Henson's Kermit, and many other iconic characters, creating personas that persist in linear time-based media. These icons indicate that the key to solving embodied agent believability problems may be discovered in simulating the performer's gestural quality. Embodied agents in games have the potential to become interactive New Media icons if we can represent the performer's craft as procedural algorithms. Our embodied agent architecture is intended as a performative component in a broader Narra-Performa-Ludic system design (see Figure 3), and is the first step in a formal Performatology study of real and simulated embodied performers interacting together in a shared performance space. Our goal is to get the professional performer in the game, and get the actor back on the stage of interactive drama.

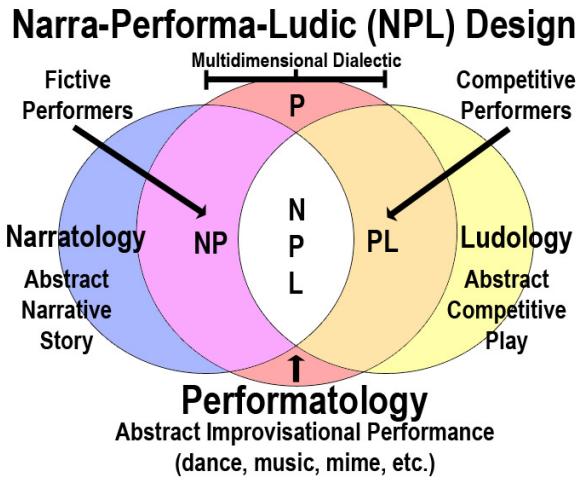


Figure 3

Acknowledgements: The author would like to thank his current Computer Science PhD advisors: Arnav Jhala (UCSC Computational Cinematics Studio) and Noah Wardrip-Fruin (UCSC Expressive Intelligence Studio). Also, a very special thanks to his Digital Arts and New Media MFA Thesis Committee: Kathy Foley (UCSC Theater Arts), Michael Mateas (UCSC Expressive Intelligence Studio), and Ted Warburton (UCSC Theater Arts). Additional thanks to his UCSC Theater Arts faculty advisors for the 2010 DANM Performative Technologies Group: Jim Bierman, David Cuthbert, and Kimberly Jannarone.

References

Aston, Elaine, Savona, G. (1991). *Theatre as Sign System: A Semiotics of Text and Performance.* New York: Routledge.

Austin, John (1962). *How to Do Things With Words*. Massachusetts: Cambridge.

Ayat-Confino, Irene (1987). *Beyond the Mask: Gordon Craig, Movement, and the Actor*. Southern Illinois Univ. Press.

Barba, Eugenio, Savarese, N. (2005). *A Dictionary of Theatre Anthropology: The Secret Art of the Performer.* New York: Routledge.

Craig, Edward Gordon (1907). *On the Actor and The Uber-Marionette, The Mask* (Florence: Self-Published Periodical.

Craig, Edward Gordon (1963). *The Theatre Advancing.* New York, NY: Benjamin Blom, Inc..

Laurel, Brenda (1991). *Computers as Theatre.* New York, NY: Addison-Weseley.

Magerko, B, Fiesler, C, et al. (2010). '"Bottoms Up: Improvisational Mico-Agents"'. *3rd Workshop on Intelligent Narrative Technologies* , FDG (INT3 2010).

Maraffi, Christopher (2010). *Avatar Theatre MFA Thesis Performances.* UCSC Digital Arts and New Media Program. `http://www.chrismaraffi.com`.

Maraffi, Christopher, Jhala, A. (2011). *Computational Cinematics Studio Performatology Project: IMPRSONA Performative Agent Architecture.* UCSC Computer Science Program. `http://www.performatology.com`.

Mateas, Michael (2004). '"A Preliminary Poetics for Interactive Drama and Games"'. *First Person* Wardrip-Fruin and Harrigan, eds.Cambridge, Massachusetts: The MIT Press.

Murray, Janet (1998). *Hamlet on the Holodeck: The Future of Narrative in Cyberspace.* Massachusetts: The MIT Press.

Neff, M, Kipp, M, et al. (2008). 'Gesture modeling and animation based on a probabilistic re-creation of speaker style'. *ACM Transactions on Graphics.* V. 27.

Perlin, Ken, Goldberg, A. (1996). '"Improv: A System for Scripting Interactive Actors in Virtual Worlds"'. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques.* New Orleans.

Schechner, Richard (2002). *Performance Studies: An Introduction.* New York: Routledge.

Searle, John (1969). *Speech Acts.* New York: Cambridge University Press.

Seif El-Nasr, Magy. (2004). 'A user-centric adaptive story architecture: borrowing from acting theories'. *ACE '04 Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*.

Tanenbaum, J., Tanenbaum (2008). '"Improvisation and Performance as Models for Interacting with Stories"'. *ICIDS '08 Proceedings of the 1st Joint International Conference on Interactive Digital Storytelling: Interactive Storytelling.* .

Taxidou, Olga (1998). *The Mask: A Periodical Performance by Edward Gordon Craig.* UK: University of Edinburgh.

Thomas, Frank, Johnston, O. (1981). *Disney Animation: The Illusion of Life.* New York: Disney Editions.

Good Evidence is Hard to Find: Policy-based Approaches to Curating and Preserving Digital Humanities Data

Marciano, Richard
richard_marciano@unc.edu
University of North Carolina Chapel Hill USA

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

Chassanoff, Alexandra
achass@email.unc.edu
University of North Carolina Chapel Hill USA

Aschenbrenner, Andreas

Hasan, Adil
English Department, University of Liverpool, UK

Blanke, Tobias

What does it mean to support scholarship in the 21st century? For the digital humanist, scholarly needs range from providing long-term online access to digitized collections to developing reusable tools to discern historical patterns among large data corpuses. The so-called “age of evidentiary abundance”[1] yielded by linked data and other technological advancements presents further interpretative challenges: materials aggregated across collections need to merge content and context in a seamless research environment. Supporting all of these endeavors over the long term requires an efficient underlying structure to manage evolving technologies, scholarship needs, and research requirements. This infrastructure must ensure longevity and usability by attending to curation and preservation components. Building sustainable research environments for the long term should be seen as a core requirement of scholarly infrastructure in the digital age.

In response to the Digital Humanities 2011 theme of “Big Tent Digital Humanities”, this poster will demonstrate policy-based approaches to building sustainable infrastructures for digital research environments. Drawing on our work from key funded projects in the United States and Europe, including Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN), DCAPE, and Policy-Driven Repository Interoperability (PoDRI), we will showcase how policy- based frameworks can be applied to both small-scale institutional collections and to larger, federated multi-repository research environments. Our work provides a model for how policy- driven curation and preservation policies can be used to effectively manage digital humanities data.

Notes

1. O’Mally, Michel. "Evidence and Scarcity." [Weblog entry.] The Aporetic. George Mason University. 02 Oct 2010. (http://theaporetic.com/?p=176). 30 Oct 2010.

A Visual Interface for Exploring Language Use in Slave Narratives

Muralidharan, Aditi
aditi@cs.berkeley.edu
Department of Computer Science UC Berkeley

1. Abstract

The increasing prevalence of digitized source material in the humanities has led to uncertainty about how this suddenly available information will change scholars' research methods. What balance will scholars strike between in-depth examination of a few sources, and a more "distant reading" (Moretti 2005) of a large number of them? Our focus is specifically on text collections: comparing texts, and identifying and tracing patterns of language use. These tasks are not widely supported by any current software, but if humanities researchers want to use digitized text collections on a larger scale, they will need to do exactly such things.

We restrict ourselves to a particular collection: the North American antebellum slave narratives, written by fugitive slaves in the decades before the Civil War with the support of abolitionist sponsors. Scholars agree about the slave narrative's most basic conventions but it is likely that these narratives, with their extreme repetitiveness, may also manifest other regular features that have yet to be detected by scholars. This project aims to assist literary scholars in uncovering these patterns with computational techniques.

In collaboration with English scholars, we have built WordSeer (http://bebop.berkeley.edu/wordseer), a system that can compare two or more narratives' grammatical features, and analyze the distribution of textual patterns throughout an entire collection. Our goal is for English scholars to be able to use our system to gather accurate information about language use patterns in a way that is intuitive, and natural to them.

We will present the system currently under development, and share the lessons we have learned while building a text exploration interface for use in the humanities.

2. System Description: Computation

The goal of the computation is to extract grammatical structure from text. We use established techniques in computational linguistics that allow fast and reliable extraction of phrase boundaries, inter-word relationship categories (e.g. subject, verb, object, modifier etc.) and parts of speech. For example, the sentence, "Marsupial mammals have pouches." contains the meaningful unit "marsupial mammals" that functions as a noun and is therefore a noun-phrase. It further contains the word "marsupial" which is an adjective modifier of the word "mammals", which is a noun. The process is completely general, and can be applied to any TEI-encoded text collection.

3. System Description: Interface

Based on the features that our literary scholar collaborators found useful, our current interface supports three tasks: searching for grammatical or word-based patterns in text (Figures 1 and 2), visualizing their distribution in the entire collection (Figure 3), annotating sections of text with tags and notes (Figure 4).

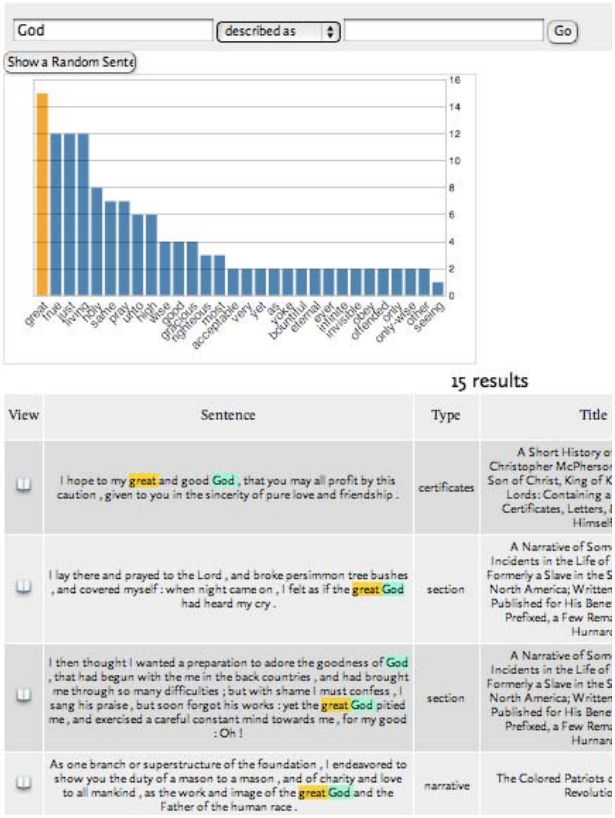


Figure 1: Easily expressing the query, "How was God described?"

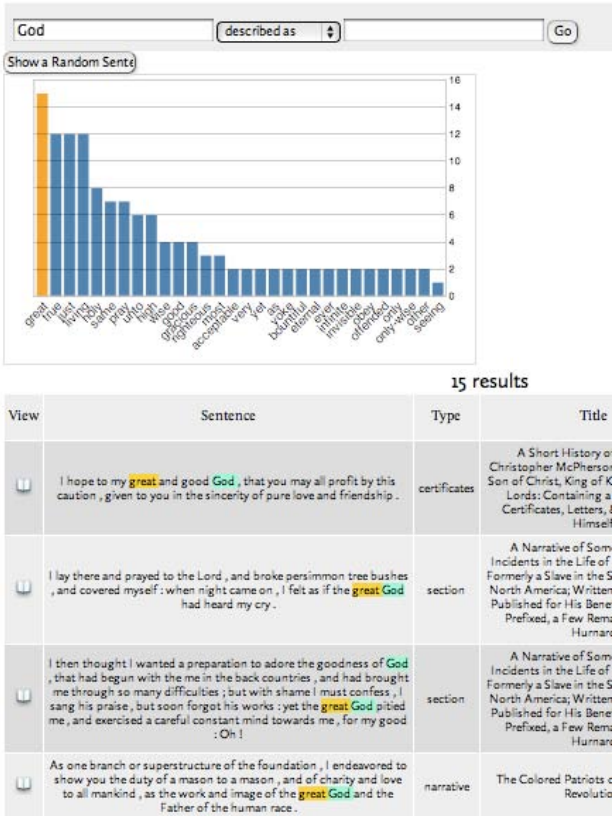


Figure 2: A list of search results augmented with a frequency visualization.

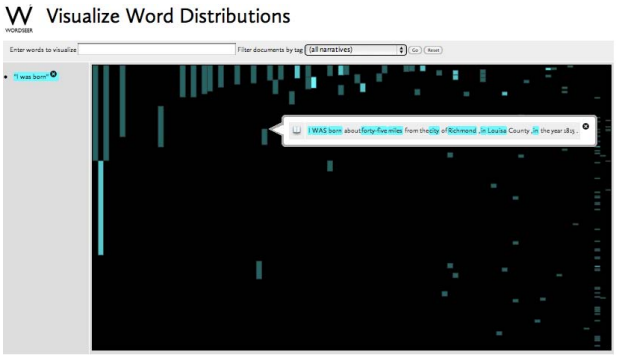


Figure 3: The distribution of the exact phrase "I was born" through the collection.

The power of extracting grammatical information is that a list search results is no longer an opaque list: trends and comparative frequencies can be extracted and displayed, giving an instant high-level picture: a guide to further exploration. For example, if a researcher were interested in the relationship between slaves and God, he or she might be interested in how God was described in the collection. Grammatical search makes this query easy to express: just type in "God" and choose the "described as" relationship, as shown in Figure 1

The search results are augmented with the simple but powerful visualization of relative frequencies shown in Figure 2: all the words that God is "described as" arranged from most to least frequent. The presence of the adjectives "great, true, just" immediately evokes a picture of the relationship - one very different from the picture more negative adjectives might paint. While this is no substitute for careful literary analysis, it can be a quick way to judge the extent to which an entity or event is represented a certain way in a collection, and so help formulate new hypotheses.

While investigating stylistic similarities between documents in a collection, it is useful to be able to investigate occurrences of patterns of interest and compare their distributions across documents. We use a visualization called heat maps, which uses the visual metaphor of text as a brick wall, with each brick, a section of text, and each column of bricks a document. Typing in a phrase shows its distribution throughout the entire collection, and patterns, such as the overwhelming occurrence (Olney 1984) of the exact phrase "I was born" at the very beginnings of narratives (Figure 3) are easily apparent.

The third research behavior we support is that of organizing sections of text into sets that illustrate a point. With our reading and annotation interface (Figure 4), researchers can highlight sections of text, and add tags and detailed notes.

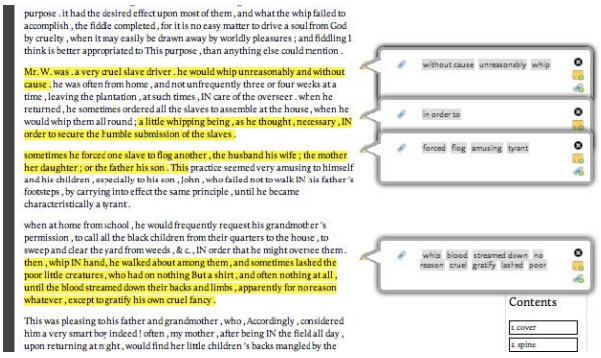


Figure 4: Reading and annotating documents with tags and notes.

Adding a tag to a highlighted section of text is like adding it to a set. Researchers can use the sets they make in other parts of the application: searching within a set, or restricting the heat map visualization to documents in a set.

4. Related Work

In the digital humanities, the closest work to our project comes from two well- known text analytics efforts. The first is the MONK project (MONK) incorporating the SEASR analysis toolkit. These projects offers

two computational linguistics tools in addition to word distribution and frequency statistics: tagging words with their parts of speech and extracting named entities. Users can visualize occurrence patterns of word sequences within a chosen text, and plot networks of how often named entities occur near each other. This research led to visual text-mining analyses of Emily Dickinson's correspondence (Catherine Plaisant et al. 2006), and of Gertrude Stein's "The Making of Americans" (Don et al. 2007) and an interface for exploring the parts of speech used near query words of interest (Vuillemot et al. 2009).

The second is Voyeur (Voyeur), which operates entirely at the word level. It allows users to plot word frequencies, see concordances (contexts in which words occur) and create tag clouds.

Other digital humanities projects have used more advanced language processing, but have not developed them into user interfaces or combined them with visualizations. Topic modeling is being applied to 19th Century British and American novels (Jockers 2010). These novels were also the subjects of cutting-edge computational linguistics research that showed how to automatically extract social networks from free text (Elson et al. 2010). Topic modeling is also being applied to the compendium of Danish, Norwegian, and Swedish folklore collected by Evald Tang Kristensen. In the field of visualization, applications to text in the humanities have been limited to word clouds, and node-and-link diagrams of named entities, and co-occurrences.

Outside the realm of text, but in the domain of comparative exploration, LISA, a comparison search interface for cultural heritage artifacts was created by (Amin et al. 2010).

The digital humanities work described above comes from the application of ideas from human-computer interaction and natural language processing. We are informed by general principles of search user interface design described by Hearst (Hearst 2009), and of visual exploration of large data-sets described by Shneiderman (Shneiderman 1996).

References

Amin, A.K. et al. (2010). 'Designing a thesaurus-based comparison search interface for linked cultural heritage sources'. *Proceeding of the 14th international conference on Intelligent user interfaces.* Pp. 249-258.

Voyeur Tools: See Through Your Texts | Hermeneuti.ca - The Rhetoric of Text Analysis. `http://hermeneuti.ca/voyeur` (accessed October 29, 2010).

Don, A. et al. (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.* Lisbon, Portugal, pp. 213-222.

Elson, D. K., Dames, N., McKeown, K. R. (2010). 'Extracting social networks from literary fiction'. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* Pp. 138-147.

Hearst, M. (2009). *Search user interfaces.* http://searchuserinterfaces.com: Cambridge Univ Press.

Jockers, M L.. *What is a Literature Lab: Not Grunts and Dullards | Matthew L. Jockers..* `http://https://www.stanford.edu/~mjockers/cgi-bin/drupal/node/45` (accessed October 29, 2010).

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract models for a literary history.* Verso Books.

Olney, J. (1984). '"I Was Born": Slave Narratives, Their Status as Autobiography and as Literature'. *Callaloo.* 46-73.

Plaisant, C, et al. (2006). 'Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces'. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries.* Chapel Hill, NC, pp. 141-150. `http://portal.acm.org/citation.cfm?id=1141753.1141781&coll=GUIDE&dl=GUIDE&CFID=110891787&CFTOKEN=59289750`.

Shneiderman, B. (1996). *The Eyes Have it: A Task by Data Type Taxonomy for Information Visualizations.* `http://ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts- Bibliography/96-13html/96-13.html` (accessed April 22, 2010).

Vuillemot, R. et al. (2009). 'What's being said near "Martha"? Exploring name entities in literary text collections'. *In Visual Analytics Science and Technology: VAST 2009. IEEE Symposium.* 107-114.

Toward a Digital Research Environment for Buddhist Studies

Nagasaki, Kiyonori
nagasaki@dhii.jp
International Institute for Digital Humanities

Tomabechi, Toru
tomabechi@dhii.jp
International Institute for Digital Humanities

Shimoda, Masahiro
shimoda@l.u-tokyo.ac.jp
University of Tokyo

In recent times, digital resources have taken on steadily greater importance in the field of Buddhist studies, with increasing numbers of digitized versions of Buddhist canonical texts, representation of material culture, and other objects of research becoming available on Web. However, despite the basic availability of such resources, most of them are not set up in an optimal way for usage by researchers; nor are they for the most part integrated with each other. For example, there does not yet exist a system that can operate with equal efficacy with philological data related to Sanskrit, Chinese, Tibetan language materials. Therefore, a comprehensive and concrete framework is needed. Although a method of text description is fairly well established in the form of TEI P5, neither the interfaces, nor the methods of presentation of results for digitized works are as yet satisfactory for the scholars of Buddhism. In this paper, we will present our approach to the establishment of requirements for various kinds of materials used in Buddhist studies and make some suggestions for the implementation of more functional interfaces as a Web research environment for such scholars.

At first, it must be noted that it is very difficult to define adequate requirements for the full range of the scholars of Buddhism, who come from a broad array of language training and methodological approaches. Thus, this paper will focus primarily on the fulfillment of the requirements for the scholars who are dealing with authentic scholarly digitized texts. In the field of Buddhist studies, where texts have been translated across a number of languages in different regions at various points in history, we have no recourse but to deal with several versions of a text, including transmitted, diffused, and translated variations at the same time (Fig.1).

(Fig. 1. An Example: " मूलमध्यमककारिका (Madhyamakakārikā)" and some of the related texts)

As discussed by Steinkellner (1988), it is often difficult to "recover" the original form of any given text, as they have been changed variously in their long tradition. In many cases, all that has come down to us is a translation that was preserved in the Chinese tradition since the 2nd century or in the Tibetan tradition since the 8th century. It is quite often the case that various witnesses are extant in both traditions. In such cases, various diplomatic texts in various languages must be compared in various units such as at the level of text, chapter, fascicle, sentence, word, syllable, or even character. Therefore, it is necessary for textual scholars to prepare an environment that delivers integrated views of a given text views. On the other hand, it is important to understand the background thought and beliefs reflected in each stage of the textual development—not only in Sanskrit and Pāli which are the closest to the original form—but also in Tibetan, Chinese. In this case, each annotation must be recorded in the context of its diffusion into other texts. In addition, modern translations of each text should be pointed to from such texts. Therefore, it is crucial for Buddhist textual studies to provide such information within a system of intertextual relationships. (See Fig. 1)

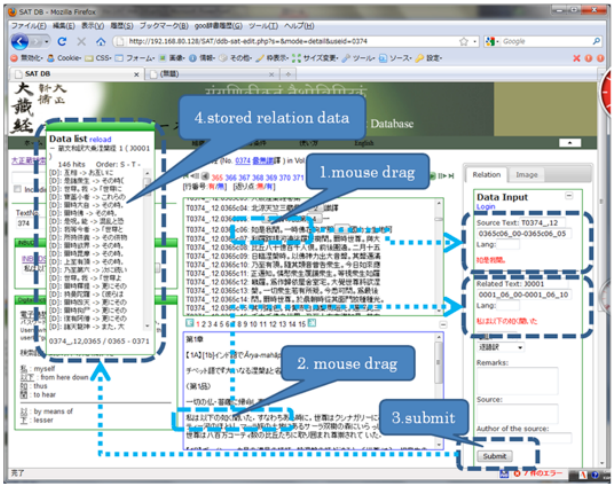
In order to realize such relationship in digitized resources, it is necessary to provide the data in a punctuation-neutral manner. This is because separation of words or sentences is not always obvious; punctuating text itself amounts to an independent and original scholarly contribution, especially in the cases of Sanskrit and classical Chinese manuscripts. Therefore, the basic concepts are:

1. being based on a unit that is not restricted to the legacy media.

2. inheritance of the legacy studies on the paper media.
3. DB providers prepare the space for the sharing of data units.
4. the users act as recipients of the units and some of them act as distributors of the units.
5. DB providers develop and distribute their own Web API so that users can make arbitrary links to access each others' data.
6. The above are realized as a collaborative research environment on the Web.

In addition, in order to preserve compatibility with past research results, all units should be identifiable in legacy media through traditional referencing methods such as T0001,01,0001a01 which means *Taishō Daizōkyō*, vol.1, page 1, register a (among a, b, and c), line 1. Then, they should be located by URI by implementation of Web API so that the other persons or applications can freely refer to arbitrary units through Web.

In order to realize such concepts, a collaborative editing system needs to be developed. Initially, users should be registered according to their own roles such as visitor, editor, or administrative editor. The role of the editor is that of inputting and checking of the data; the role of administrative editor is checking of the data and the determination of the distribution of the data. Both editor and administrative editor should be able to efficiently select arbitrary units with an easy method such as mouse click or drag on the text body without any restriction posed by the textual structure. Then they can append any necessary information to the units and link selected units to other ones (Fig. 2).



(Fig. 2. The procedures of inputting relationship data on an environment)

In the final step, the administrative editor should check and determine whether each unit is ready for release. The role of visitor is only to browse the workspace to observe the progress of the work. On the other hand, anonymous users can browse only released data. They can efficiently view various information with various methods such as mouse click or drag on the text body and select (or ignore) any data on the basis of various properties such as their sources, contributors, and so on. In order to realize such a function on Web, AJAX will be fully utilized.

Finally, we will explain the Web Database function. The Web Database stores only relational data that includes one or two locations in the original materials, its annotation or relationship information, contributor's name, data composition and other related information. The writing rule of the location allows the user to use the information even without digitized textual material. However, if the material is originally in digitized form, the data must refer to a logical location such as the URI. Users don't need to be aware of separation between the database and the materials presented to them because to the viewer of the materials, they seem to be seamlessly integrated with the database. By this kind of method, it will be easier to integrate the other materials which are released on other Web sites into the environment. The database can provide arbitrary data according to the user's preference through Web API, as well as retrieve and show the data provided by other DBs. From here, the data can be published under various formats such as RDF. However, in principle, this kind of system would not be able to avoid the problem of overlap if there is an intent to publish its data fully including the original materials. So we are trying to do so according to TEI P5 by using a kind of stand-off markup.

(Fig. 3. The procedures of browsing the relational data on an environment)

Although we have just begun to develop this approach, it has already been tested by some scholars and we have seen positive results. Actually, these early testers say that it is very useful for them because their accessibility and permissions are limited and explicitly shown. However, some enhancements have been suggested, which mainly concern the function of browsing. We will work on this until the time of DH2011, when we will able to present a much more matured system.

References

Burnard, L. , Bauman, S. (2007). *P5: Guidelines for Electronic Text Encoding and Interchange.* http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Caton, P. (2007). 'Distributed Multivalent Encoding'. *Digital Humanities 2007.* University of Illinois, Urbana-Champaign, 2-8 June, 2007, pp. 33-34.

DeRose, S. (2004). 'Overlap: A Review and a Horse'. *Extreme Markup Languages 2004.* Montreal, 2-6 Aug., 2004. http://www.mulberrytech.com/Extreme/Proceedings/html/2004/DeRose01/EML2004DeRose01.html.

Lavagnino, J. (2009). 'Access'. *Literary and Linguistic Computing.* 24 (1): 63-76.

Nagasaki, K. (2008). 'A Collaboration System for the Philology of the Buddhist Study '. *Digital Humanities 2008.* Oulu, Finland, 25-29 June, 2008, pp. 262-263.

Rehm, G. , Witt, A. (2008). 'Aspects of Sustainability in Digital Humanities '. *Digital Humanities 2008.* Oulu, Finland, 25-29 June, 2008, pp. 21-29.

Renear, A. H. (2004). 'Text Encoding'. *A Companion to Digital Humanities.* Schreibman, S., Siemens, R., Unsworth, J. (eds.). Oxford: Blackwell, pp. 218-239.

Steinkellner, E. (1988). 'Methodological Remarks On The Constitution Of Sanskrit Texts From The Buddhist Pramāṇa-Tradition'. *Wiener Zeitschrift für die Kunde Südasiens, Band XXXII.* Pp. 103-129.

ArchiTrace: An Urban Social History and Mapping Platform

Nieves, Angel David
anieves@hamilton.edu
Hamilton College, United States of America

ArchiTrace is a dynamic, web-based markup tool that will allow researchers and content authors to work together in a collaborative co-authoring environment as they build and share architectural drawings, maps, and spatial representations of African urban spaces across a historical timeline. By focusing on the spatial dimensions of the built environment, historical and present-day cityscapes become both the backdrop and interface to a rich archive of cultural heritage assets gathered over the course of their historical development. *ArchiTrace* will combine a broad range of artifacts such as historical texts and photographs, maps and aerial photographs, providing both cultural and historical context to the depicted regions. Finally, by allowing an audience to watch the transformation of both the spaces and their cultural artifacts evolve over expansive timelines, researchers will be able to follow the historical tracks of the evolution of the regions themselves.

In the example of Soweto, the model housing schemes of the 1950s through the 1980s provided by the state government and its architects were essentially an attempt at social control, but through the inspection of aerial photography and the differences in what was proposed, built, and modified by residents reveals a certain amount of resistance to the ordered panopticon of township housing schemes.

Through the aerial photography captured over a 50-year period, the changes in the city's physical landscape suggest a form of spatial resistance to the imposed uniformity of the built environment under apartheid. By facilitating architectural and spatial comparisons of this kind through a suite of interactive mapping and comparative tools, *ArchiTrace* will allow historically embedded social histories to become emergent features of the overall presentation. Where projects like HyperCities focus on the development of the urban form, *ArchiTrace* facilitates the curation of a biographical narrative in a city's social history as a way to characterize the development of a region and its people.[1] HyperCities provides a blank canvas for understanding urban form, while *ArchiTrace* will provide a spatial subtext to urban place making.

ArchiTrace's granular data structure will allow users to inspect fine detail in a city environment at any level of scope, whether tracking the development of entire neighborhoods over decades, or the changes of an individual building's architecture over the course of mere years. *ArchiTrace* will offer content authors tools for the creation of data units at several different scopes. Content authors can "curate" the granular layering of spatial information from the level of individual home, block, town or city, to more generalized regional levels. *ArchiTrace's* uniquely granular data structure is inspired by architectural "Building lives:" mapping the life cycle of an urban form across a variety of scales and time periods.

ArchiTrace will feature prominent social networking and collaborating tools, allowing researchers and content authors the opportunity to build, share, and co-edit materials within a project collection. This can range from the simple sharing of files within a project's workspace to the realtime shared editing of text, architectural drawings, and annotations. *ArchiTrace* aims to facilitate researchers and participants alike in the creation of both collections and exhibits, acting as both an archival and educational platform for researchers and public audience.

By engaging open source software and standards, *ArchiTrace* aims to create an extensible project platform capable of growing in the many directions promised by forthcoming web technologies. As web-based 3D technology becomes more broadly implemented in future web browsers, we anticipate being able to integrate the models from which our images are derived into an interactive and dynamically-generated 3D environment. Although the initial depictions of *ArchiTrace's* cityscapes will begin as 2-dimensional representations, we aim to eventually extend those aerial photographs, maps, and still-image renderings of the project into 3D models.

A recent series of discoveries, in the South African National Archives, provides for some new ways to begin mapping the growth and development of townships over the forty-six year apartheid era. Over the past four months I have come across a cache of maps, architectural plans, aerial photographs and other source documents related to the design and planning of townships across the city of Johannesburg. These drawings and plans suggest, as many historians have previously stated, "the apartheid state remained steadfastly committed to terror" through the design and "layout of the location [townships which] were planned with explicit, detailed attention to the disciplinary potentials of space."[2] As some have argued, "The tyranny of the planners' blueprints yielded a degree of spatial compartmentalization whose sheer banality had profound implications for every aspect of urban life."[3] As noted, "when planners reshape the built environment, individuals are compelled to adjust accordingly, reinforcing to some extent the spatial parameters of their oppression."[4] I am however suggesting that individuals and communities impart changing meanings to spatial structures over time. Space not only becomes implicated in the transformation of self-perceptions and the capacities of social groups – space both restrains and enables in some deeply profound ways. Township residents "were neither simple victims of the state nor pure protagonists of resistance" – they negotiated their daily lives through their engagements with racialized space-making. Resistance, among those incarcerated/detained/imprisoned in "planned communities" like Japanese Americans in internment camps, or victims of the Holocaust through concentration camps, is as much about the forging of bonds within space as it is about propelling struggles from one "stage" (or township), to another. How then might we document the "spatial underpinnings of apartheid [era] projects"? After the Soweto Uprising of 1976, the spatial compression, brought on by township planning and design, enabled for the mass proliferation of dense political networks within and between Black residential areas across Johannesburg. How might we then use spatial tools to document the history of resistance, while also telling the story of the ways in which residents converted the bureaucratic and spatial impediments to political mobilization into weapons of struggle? These and other questions – that have yet to be formulated – are the basis for developing ArchiTrace.

Notes

1. Todd Presner, "Hypercities: A Case Study for the Future of Scholarly Publishing," The Shape of Things to Come, ed. Jerome McGann (Houston: Rice University Press, 2010), 251-71.
2. Ivan Evans, Bureaucracy and Race: Native Administration in South Africa (Berkeley: University of California Press, 1997).
3. Ibid.
4. Ibid.

An Analysis of Recurrences in Harold Pinter's Plays Using CATMA Concordancing Software

Onic, Tomaz
tomaz.onic@uni-mb.si
Department of English and American Studies,
University of Maribor, Slovenia

Recurrence is a crucial feature contributing to a recognizable style of dramatic characters in the plays by contemporary British playwright Harold Pinter. His characters sometimes repeat whole phrases or sentences, sometimes with slight changes, which almost always indicate a change in the speaker's intention. The repeated passage either follows its first appearance closely, or can be delayed for a few lines – or sometimes pages. In general, the most noticeable recurrences for the audience are those consisting of multiple repetitions, containing unusual words or phrases attracting our attention, or consisting of closely repeated passages.

Beaugrande and Dressler (1988, 54) define *recurrence* as a direct repetition of a textual element which has appeared earlier in the text. They do not call it repetition because, according to Beaugrande (1991, 18), only seldom is the repetition of part of text a real repetition. Such *absolute recurrence*, as he calls it, would have to carry exactly the same meaning potential of the repeated phrase as did its first appearance. In most cases, that does not happen, as it is usually the very intention of the speaker that causes the recurrence: "Saying the same thing over again normally carries a context-sensitive message, such as approval, insistence, anxiety, doubt, surprise, or irony. /…/ thus, recurrence is typically an instance of 'incremental recursion', where the repeated event adds to the value of the original" (Beaugrande 1991,18).

A less strict variation of recurrence is *partial recurrence*, defined by Beaugrande and Dressler (1988, 54-55) as the re-appearance of a certain word in the form of a different part of speech. As such, it is similar to polyptoton, a figure of speech that is often defined as repetition of the same word in various inflected forms. In a later article, Beaugrande (1991) defines it as the repetition of a word cluster that does not repeat as a whole; not even all the elements need to repeat. It suffices that some elements of the original sentence repeat in the same or a sufficiently similar form.

Some partial recurrences (which can also originate in language system functions) are random; others appear as a result of the writer's or speaker's intention. Pinter's characters can be classified into a combined category, since their speech – together with recurrences – is carefully designed in order to sound random.

The role of recurrence in Pinter's dramatic texts represents an important translation issue. It is vital that this stylistic feature be preserved in the target language as faithfully as possible, since it is not only an important decorative device but represents one of the key Pinter's stylistic trademarks. Unfortunately, however, the existing research results have shown that this stylistic element is often disregarded in translation. It often comes second to meaning or other similar language elements, but usually its loss can be attributed to the translator's lack of awareness of the importance of recurrence.

A substantial potential danger of omitting recurrence from translation is the fact that this is not a feature that disturbs the audience with its absence. For this reason, it may seem a harmless translation shift; however, it is, in fact, damaging to the audience's perception of Pinter's style, which is skewed owing to the absence of such an important stylistic element.

Having been involved in Pinter Studies for about ten years, I have conducted several research activities concerning recurrence in Pinter's plays – into originals as well as translations – manually. Therefore, the natural next step was to perform selected parts of this body of analysis again using digital methods and thus confirm and broaden – possibly also correct – the existing results.

The first step in the project was to digitize and mark up the original texts and their Slovene translations. These texts were then uploaded into the new CATMA (Computer-Aided Textual Markup and Analysis) concordancing program. Since CATMA allows for searching not only by word or phrase, but also by "grade of similarity", it is expected that this will be useful in attempting to identify those phrases which are not exact repetitions, but which are similar enough to be construed as a type of repetition in the reader's mind. This will also lend more weight to any confirmation of the existing idea that recurrences often are not preserved in translation. Figure 1 shows how CATMA allows searching with "grade of similarity" for phrases close to "like a"; when the results are shown on the screen, you can then select an item

or not depending on whether it is similar enough to possible be what you are looking for. In this case you would definitely not select the first two examples, ie detail and Special. Since the similarity percentage requested here is only 60%, it is understandable that such items will be deemed as similar; should you raise the measure to 80%, they, and other items like them, will not be included in the results.

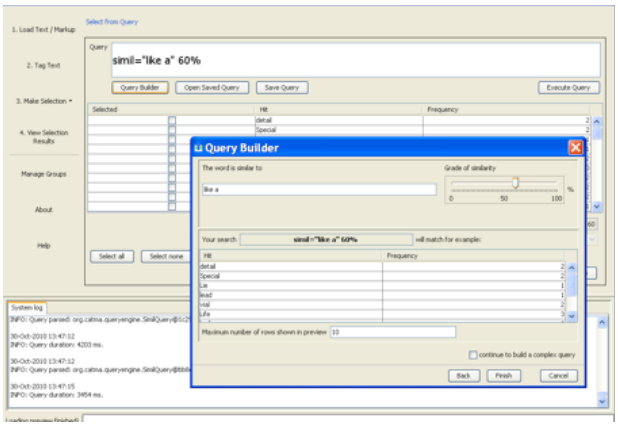


Fig. 1 Screen shot of the similarity function in CATMA

Hopefully, the results obtained through this research will strengthen the awareness of how important it is to consider this element in translation, and potentially contribute to a higher quality of the translation theory and practice in the Slovene cultural space of translated foreign language literature.

I feel that participating at the DH2011 would offer a good opportunity to discuss this undertaking and its practical implementation with other participants who have experience with similar projects, particularly in the field of drama which has its own specifics.

References

Beaugrande, R. de (1991). 'Coincidence in Translation: Glory and Misery Again'. *Target.* 1: 17-53.

Beaugrande, R. de in W. U. Dressler (1988). *Introduction to Text Linguistics.* Harlow: Longman.

University of Hamburg, Germany. *CATMA*. `http://www.slm.uni-hamburg.de/catma/index.html`.

Pinter, H. (1977). *Complete Works: Two.* New York: Grove.

Distributed Access to Oral History collections: Fitting Access Technology to the Needs of Collection Owners and Researchers

Ordelman, Roeland J.F.
ordelman@ewi.utwente.nl
University of Twente, Enschede, The Netherlands
Netherlands Institutefor Sound and Vision, Hilversum, The Netherlands

1. Introduction

In contrast with the large amounts of potential interesting research material in *digital multimedia repositories*, the opportunities to unveil the gems therein are still very limited. The Oral History project 'Verteld Verleden' (Dutch literal translation of Oral History) that is currently running in The Netherlands, focuses on improving access to *spoken testimonies* in collections, spread over many Dutch cultural heritage institutions, by deploying modern technology both concerning infrastructure and access. Key objective in the project is mapping the various specific requirements of collection owners and researchers regarding both publishing and access by means of current state-of-the-art technology. In order to demonstrate the potential, Verteld Verleden develops an Oral History portal that provides access to distributed collections. At the same time, practical step-by-step plans are provided to get to work with modern access technologies. In this way, a solid starting point for sustained access to Oral History collections can be established.

2. Technology and Daily Practice

The Verteld Verleden project builds upon years of academic research on access technology for spoken word archives [1] deploying among others automatic speech recognition, text-to-speech synchronization and fragment-level search. This research resulted in a number of demonstration applications in close cooperation with cultural heritage institutes [2]. Although these demonstrations have been very well received both by the public and involved institutes, it was observed that there is still a gap between academic, 'technology-driven' pilots and the daily practice of

Dutch cultural heritage institutions and researchers. One important observation was that Oral History collections are managed in many different ways: from very adequate to not at all. There are a few 'forerunners' that do already make use of various professional infrastructures and technologies for disclosure and access but the larger part of the collection owners, in spite of acknowledging the virtues of modern access technologies, often do not have the knowledge or means to really start using these. In practice, access to Oral History collections in general is very limited and publically accessible overviews on available collections are missing. On the other hand, we see that collection owners and researchers have very specific requirements with regard to management, disclosure and access, and have very detailed knowledge of their collections and their contexts.

3. Knowledge Transfer

In order to be able to catch up with the advantages of the digital networked society, cultural heritage institutes need practical handles that are specifically geared towards their specific use cases. The Verteld Verleden project follows this practical approach by mapping the available solutions and best practices in The Netherlands on a diversity of relevant topics ranging from digitization of audio-visual data, format conversion, online access to collections, and (semi)automatic metadata generation, and linking collections to other information sources, to dealing with privacy/copyright issues. On an academic level, special attention is addressed towards transfer of knowledge and awareness on methodology and theory of Oral History research, and the design of Oral History research in combination with modern technology.

4. Showcase

To showcase how these best-practices and solutions could work out in practice the project builds an embeddable Oral History portal that enables access to distributed oral history collections. The general approach is that collections owners are urged to comply with interoperability standards on the dissemination of metadata and content. By adopting these standards, content owners allow aggregators to channel content into local portals (e.g., Verteld Verleden) or even international portals such as the Europeana portal [3]. The Verteld Verleden portal serves here as a so called thematic portal.

Verteld Verleden promotes OAI-PMH, the Protocol for Metadata Harvesting and stimulates content owners to have their content available using a streaming media protocol to enable play-out of search results. The Verteld Verleden portal harvests metadata from associated institutes and provides centralized search for searching and browsing the collections that are linked up. As the portal's user interface can also be embedded in the local websites, content owners can be provided with search functionality for their own content.

5. State-of-the-art

The portal is equipped with state-of-the-art search technology and a flexible user interface that allows the project to adapt it easily to the requirements of researchers and content owners that are expected to advance during the project as a result of discussions at workshops and local expert sessions. An important requirement of researchers is evidently to have sophisticated means to access and analyse available Oral History collections. To a large extent however, access to collections is rather limited due to the lack of appropriate *fragment-level* semantic descriptions. Metadata is often only sparsely available, forcing scholars to play an A/V item in full in order to decide if, and if so, which parts of the material are of interest for their research. Moreover, exploring possible *correlations and connections* both *within and across large data collections* requires an additional layer on top of the metadata for the interlinking of multimedia content sources and/ or collection fragments. Ultimately, also dedicated technology for *browsing, accessing, analysing and comparing sources* effectively during the various phases of research (exploration, analysis, publication, verification) are a prerequisite for the innovation of the methodological framework of humanities researchers and for the formulation of new questions and the renewal of research agendas. In order to successfully exploit these technologies for the purpose of humanities research, their development must strongly be steered by the demands and requisites of the researchers and their research paradigms.

6. Speech Recognition

A special role in the project is assigned to the use of speech recognition technology. Speech recognition can play an important role in the process of making Oral History content better accessible, either directly via the conversion of speech to text or indirectly using available textual transcripts and a technology derived from speech recognition often referred to as forced-alignment. Verteld Verleden offers associated content owners the use of a speech recognition service

supported by the Dutch CATCHPlus program [4], that aims to valorise scientific research results to usable tools and services for the Entire Dutch heritage sector.

Deploying speech recognition brings up an additional challenge with regard to metadata models and harvesting standards: it encompasses the need to incorporate time-labelled into the metadata model. Approaches are currently investigated in close collaboration with CLARIN-NL, a project on Common Language Resources and Technology Infrastructure [5]

References

Verteld Verleden homepage: . http://www.verteldverleden.org.

CATCHPlus program. http://www.catchplus.nl/en/.

Notes

1. Goldman, J.Renals, S. Bird, S. de Jong, F. M. G. Federico, M., Fleischhauer, C. Kornbluh, M. Lamel, L. Oard, D. W., Stewart, C.Wright, R. (2005) 'Accessing the spoken word'. *Int. Journal on Digital Libraries.* 5(4)287–298 F.M.G. de Jong D.W. Oard W.F.L. Heeren R.J.F. Ordelman. 'Access to recorded interviews: A research agenda'. *ACM Journal on Computing and Cultural Heritage (JOCCH).* 1(1)3-29 (2008) R.J.F. Ordelman W.F.L. HeerenM.A.H. HuijbregtsF.M.G. de Jong D. Hiemstra. 'Towards Affordable Disclosure of Spoken Heritage Archives'. *Journal of Digital Information.* M.A. Larson, K. Fernie and J. Oomen (eds)10(6)17-33 (2009)
2. Radio Oranje Demonstrator (alignment of historical speeches): http://hmi.ewi.utwente.nl/choral/radiooranje.html Searching interviews bombarding of Rotterdam: http://www.gemeentearchief.rotterdam.nl/brandgrens/navigator/interviews.php Access to interviews with survivors of World War II concentration camp Buchenwald: http://www.buchenwald.nl
3. Europeana portal: http://www.europeana.eu/portal/
4. OAI-PMH: http://www.openarchives.org/pmh
5. CLARIN-NL: http://www.clarin.nl

A Collaborative Linguistic Research Interface for the 1641 Depositions

O'Regan, Deirdre
deirdre.oregan@gmail.com
School of Language and Literature, University of Aberdeen, King's College, UK

Sweetnam, Mark
sweetnammark@gmail.com
School of Language and Literature, University of Aberdeen, King's College, UK

Fennell, Barbara
b.a.fennell@abdn.ac.uk
School of Language and Literature, University of Aberdeen, King's College, UK

Lawless, Seamus
seamus.lawless@scss.tcd.ie
Knowledge and Data Engineering Group, Trinity College Dublin, Ireland

This poster presents an account of the development of a collaborative research environment for the socio-historical linguistic exploration of a unique seventeenth-century resource of Irish national importance and international significance. It is also an account of the process - from inception to application - of a highly interdisciplinary, unique collaboration between academia and industry, which is part of an evolving set of DH projects.

The '1641 Depositions' in Trinity College Dublin comprise some 8,000 personal statements, in which mainly Protestant men and women of all classes told of their experiences following the outbreak of the 1641 Rebellion in Ireland by the Catholic Irish. Collected by government-appointed commissioners, the witness testimony runs to approximately 20,000 pages, and constitutes the chief evidence for the sharply contested allegation that the rebellion began with a general massacre of Protestant settlers. As a result, this material has been central to protracted and bitter historical dispute.

This body of material, unparalleled elsewhere in Early Modern Europe, provides a unique source of information for the causes and events surrounding the 1641 Rebellion and for the social, economic, cultural, religious, political and linguistic history of

seventeenth-century Ireland, England and Scotland. In addition, the depositions vividly document various colonial and 'civilizing' processes, including the spread of Protestantism in one of the remotest regions of the Stuart kingdoms and the introduction of lowland agricultural and commercial practices, together with the native response to these developments.

Following the recent completion of a three year process of digitizing, transcribing and annotating the 1641 Depositions, the resulting Text Encoding Initiative (TEI) encoded corpus has become available for digital enhancement and analysis. The Arts and Humanities Research Council (AHRC) of the United Kingdom has funded the next generation of research on this corpus under the auspices of their 'Digital Equipment and Database Enhancement for Impact' programme. 'Language and Linguistic Evidence in the 1641 Depositions' is a multi-disciplinary Digital Humanities project designed to create an interactive computer environment in which scholars interested in historical linguistics, corpus analysis and forensic linguistics / discourse analysis can work together with historians, Early Modern prose scholars and other specialists to interrogate these valuable resources, exploiting new methods of personalization, visualization and collaboration.

The 1641 Collaborative Linguistic Research and Learning Environment (CLRLE) has been developed using Omeka, an open-source digital archival collections management system that has become a popular tool in the Digital Humanities for archiving, publishing and managing access to primary source materials such as documents, images, transcriptions and other multimedia resources (see http://omeka.org/). Omeka is an ideal tool for exploring the concept of a collaborative research interface, since it doubles as a content management system and offers myriad possibilities for personalization and collaboration amongst users.

The resulting web-based portal houses fully searchable records of the 1641 Depositions as 'Items' in various 'Collections', as is typical in Omeka-powered applications. Privileged users of the interface can collaboratively manage the archive and its content, editing Collections and Item metadata (e.g. deposition transcriptions, dates and deponent and commissioner names), annotating and tagging Items and contributing specialist content to public web pages on the site.

A central part of this project has been knowledge exchange with IBM's LanguageWare Research and Development Team. LanguageWare (http://www.alphaworks.ibm.com/tech/lrw) is IBM's natural language processing software and is part of the Unstructured Information Management Architecture (UIMA) framework (http://uima.apache.org/). Researchers have addressed the challenge of applying this software (designed for contemporary language analysis) to the highly problematic "dirty data" of the 1641 corpus, with its propensity to variable spellings, morphologic instability and syntactic complexity. This has allowed the identification of important processes of linguistic change and has enabled linguists to trace the development of English in this unique Early Modern corpus. This involved the integration of a suite of software creating a domain-specific UIMA pipeline which offered a level of accuracy comparable to that achieved by manual annotators (Sweetnam and Fennell, 2010). A crucial element of CLRLE will be the integration and exploitation of the results of this analysis.

A particularly valuable feature of the Omeka-powered CLRLE is the provision of an interactive Exhibit Builder tool enabling users to create personalized 'Exhibits' of their research outcomes. These Exhibits draw together a highly extensible collection of reusable research objects, including transcribed depositions and associated metadata, dynamic visualizations, the outputs of statistical linguistic analyses and GIS displays. These Exhibits facilitate a high level of research cooperation and dissemination, and also have considerable pedagogical and outreach applications.

This poster charts the successful completion of a multi-disciplinary collaboration involving the adaptation and modification of new and evolving open source technologies for humanities research, significant knowledge exchange between industry and academia, and the interaction of a range of private and public institutions including the University of Aberdeen, Trinity College Dublin, Lancaster University, IBM LanguageWare and the Irish Digital Humanities Observatory (http://dho.ie/). The outcomes of this project offer valuable lessons for future undertakings in the Digital Humanities. CLRLE is an exemplar of the potential for the impact of modern technology and new methodologies on our understanding of historical resources and underlying processes and their continuing contemporary relevance.

References

Sweetnam, Mark , Fennell, Barbara (2010). 'Natural Language Processing and Early Modern Dirty Data'.

Proceedings of the Chicago Colloquium on Digital Humanities.

Trinity College Dublin. *1641 Depositions Project.* http://1641.tcd.ie/ (accessed 15 Mar 2011).

TEI Consortium, (ed.). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* http://www.tei-c.org/Guidelines/P5/ (accessed 30 Aug 2010).

Center for History and New Media (CHNM), George Mason University. *Omeka. Version 1.2.1.* http://omeka.org/ (accessed 30 Aug 2010).

Modelling a Web Based Editing Environment for Critical Editions

Litta Modignani Picozzi, Eleonora
eleonora.litta@kcl.ac.uk
Department of Digital Humanities, King's College London

Noël, Geoffroy
geoffroy.noel@kcl.ac.uk
Department of Digital Humanities, King's College London

Pierazzo, Elena
elena.pierazzo@kcl.ac.uk
Department of Digital Humanities, King's College London

Critical editions can be complex objects to digitise both in the output/publication process and in the production.

In most cases, for the output, there is the need to manage facsimiles, a main text (with lots of diacritics, deletions, suppressions, supplied text), an apparatus (with a very dense, conventional notation), footnotes (with cross references and bibliographic references), and conventional markings for page breaks from previous editions and/or manuscripts, in the form of other digitised text or images.

On the production side we have witnesses of the text, forming the direct tradition, in some cases followed by re-elaborations, contemporary or later translations, quotations that form the indirect tradition. This picture is usually completed by previous critical editions, commentaries, and a possibly large set of secondary references from other texts, dictionaries, biographies and prosopographical data.

The present paper will outline an attempt made at the Department for Digital Humanities, King's College London[1] to produce a model for new editions and translations of all English legal codes, edicts, and treatises produced up to the time of Magna Carta, 1215.

This particular project, called Early English Laws[2], presents a situation which is even more complex with respect to what has been outlined above. This is because of two characteristics typical of legal texts:

1. All the witnesses of a tradition were effective binding laws in a specific place and time, including the

mistakes added in the production of the manuscript copy.

2. Laws were constantly re-elaborated to become new laws and each of these stages of elaboration represents a fully effective law.

This means that the legal text is regularly transmitted in a number of versions, witnesses of both the direct and the indirect tradition.

As a result, each law can be represented as a universe of texts and images that relate to each other: this universe can be exemplified in the figure below, a diagrammatic representation of the so-called Treaty of Edward and Guthrum and its relations, one of which includes an Old English translation which has its own universe of relations (see figure 1)

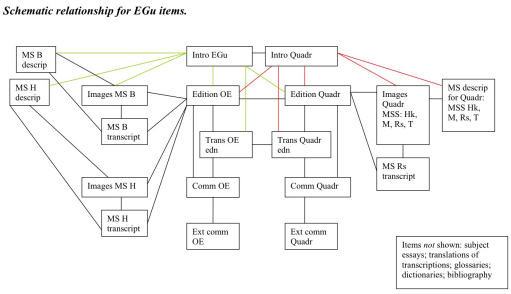


Figure 1

The development of the project presents a series of problems, for which a series of solutions are being tested and implemented. The Early English Laws project represents a case study of a more general issue, namely how to model critical edition from a conceptual point of view.

The main concern until now has been the need to create a conceptual model for the editions in order to identify the best structure to support the variety of content offered by the project. The difficulty here lies in trying to apply the philological terminology and needs to the level of conceptual discrimination required by the normalisation process involved in the design of a relational database schema that can be able to connect related law codes. For instance the term 'text' is extremely ambiguous as it actually implicitly conveys different concepts such as a work, an edition, a version, a transcription or a translation (Carlyle, 2006).

Failing to identify and address this issue would inevitably lead to a poor organisation of data. This, in turn, would confuse the editorial work and also generate inconsistencies, which will cause problems when developing the editorial environment and web site.

The solution to this problem is identifying each component of the text/image universe as entities, their properties and interrelationships can help to disambiguate concepts through the systematic analysis of the sample material. Borrowing elements from the FRBR model is also proving to be very useful in our attempt to discover and separate off hidden conceptual layers in our first version of the data model. The Entity-Relationship representation of the overall data model and the formal and careful definition of problematic concepts provides a shared vocabulary which helps the members of the project team to communicate with precision about the subject domain, whether they are from a programming or indeed from a solely humanistic background (Pierazzo, 2010). Moreover this conceptual data model serves as a basis for the transformation into a database schema, which will support the entire set of resources.

Given the variety of content involved in this project, this transformation into a database will be challenging as well. Indeed, the information captured by this project can be divided in two categories depending the format of its representation (see figure 2).

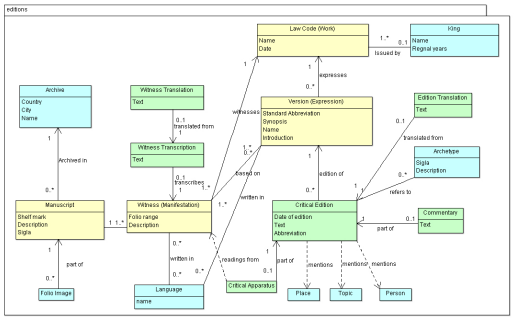


Figure 2

1. The entities identified during the modelling process, their properties and the relationships among those entities. These are ideally stored in a relational database. The relational database is also very advantageous for the storage, editing, indexing and referencing of authority lists.

2. All the information found in the text of the manuscript transcriptions or the peripheral textual information such as the critical apparatus, the introduction text for the manuscript, the comments. The best format

for this type of information is XML with conformance to the TEI guidelines.

The main challenge here is how can we harmoniously combine the two very different types of information within the same database management system, especially when we consider how deeply interrelated they are (e.g. an inline reference from one text to another or to an authority list will consist in inserting a primary key of relational tuples within an attribute of a TEI element).

To solve this, we are intending to exploit the XML-oriented facilities offered by recent relational database management systems to keep all the information centralised and linked up despite being hybrid in nature. Substantial custom development to the editing framework (possibly Django) is also expected in order to allow the user interface to acknowledge and seamlessly integrate the two types of data.

References

TEI Consortium (ed.). *TEI P5:. Guidelines for Electronic Text Encoding and Interchange.* 1.7.0`http://www.tei-c.org/Guidelines/P5/`.

Carlyle, Allyson (2006). 'Understanding FRBR as a conceptual model: FRBR and the bibliographic universe'. *Library Resources & Technical Services.* 50 (4): 264-273.

Danskin, Alan & Chapman, Ann (2003). 'Bibliographic records in the computer age'. *Library & Information Update.* 2 (9): 42-43.

Liebermann, Felix (ed.) (1903–1916). 'Peace of Edward and Guthrum'. *Die Gesetze der Angelsachsen.* , pp. 128–134.

Pierazzo, Elena (2010). 'Editorial Teamwork in a Digital Environment'. *Jahrbuch für Computerphilologie.* 10. `http://computerphilologie.tu-darmstadt.de/jg08/pierazzo.html`.

Notes

1. Formerly Centre for Computing in the Humanities. The name has been changed in spring 2011.
2. Early English Laws is currently in the early development stage and the first release of the resource can be found at http://www.earlyenglishlaws.ac.uk, where a list of texts to be encoded has been made available. Funding for the project will be available only for an initial period of three years, during which an interface for the functioning of the whole repository will need to be made available together with a full image database and a number of recently published and new editions. Completion is expected to take ten years, during which several scholars from around the world will volunteer to edit a list of available laws and input their work directly into a web-based interface. This should ideally allow for a minimal amount of reworking by one or two people volunteering to review submissions.

The Story of TILE: Making Modular & Reusable Tools

Porter, Dorothy (Dot)
dot.porter@gmail.com
Indiana University, United States of America

Reside, Douglas
dougreside@gmail.com
University of Maryland, United States of America

Walsh, John A.
jawalsh@indiana.edu
Indiana University, United States of America

The Text Image Linking Environment (TILE) is a collaborative project between the Maryland Institute for Technology in the Humanities (MITH), the Digital Library Program at Indiana University, and the School of Library and Information Science at Indiana University Bloomington. Since May 2009, the TILE project team has been developing through NEH Research & Development funding a web-based, modular, image markup tool for both semi-automated linking between encoded text and image of text, and image annotation. The software will be complete and ready for release in June 2011.

TILE was designed to change the way that people think about digital humanities tools. Many tools created for humanists are built within the context of a single project, focusing either on a single set of materials or on materials from a single time period, and this limits their ability to be adapted for use by other projects. The TILE project – not just the software, but the project itself – was designed to cut across subjects and materials. Because it is simple, with focused functionality, TILE will be usable by a wide variety of scholars from different areas and working with a variety of materials – illustrations and photographs as well as images of text. To help ensure this, we brought together several collaborators from different projects with different needs to provide advice and testing for our work: The Swinburne Project and Chymstry of Isaac Newton at Indiana University-Bloomington, the Homer Multitext Project at Harvard's Center for Hellenic Studies, and the Mapas Project at the University of Oregon. In addition, we were very fortunate during the life of the project to have volunteers test the software on their own materials, or to provide materials for us for testing. We also received feedback from user testing performed at the Department of Information Studies, University College London, and the School of Library and Information Science, Indiana University Bloomington.

The basic functionality of TILE is to create links between images and text that relates to that image – either annotations or transcriptions. We have paid particular attention to linking between image of text and transcription of text. These links may be made manually, but the project also includes an algorithm, written in JavaScript, for recognizing text within an image and automatically associating the coordinates with a Unicode transcription. Additionally, the tool can import and export transcriptions and links from and to a variety of metadata formats (TEI, METS, OWL) and will provide an API for developers to write mappings for additional formats. Of course, this functionality is immediately useful to a relatively limited set of editors of digital materials, but we have made modularity and extensibility primary goals of the project, and in our demonstration we hope to be able to show new recognition algorithms to, for instance, identify printer watermarks or panels in a graphic novel.

Many members of the TILE development team are also members of the Open Annotation Collaboration (OAC), and have therefore attempted to develop TILE's annotation features to be OAC compliant. Like OAC, TILE assumes that the text and the images to be linked may exist at separate and completely unconnected servers. When a user starts the TILE tool for the first time, she is prompted to supply a URI to a TILE compliant JSON file [pictured in figure A].

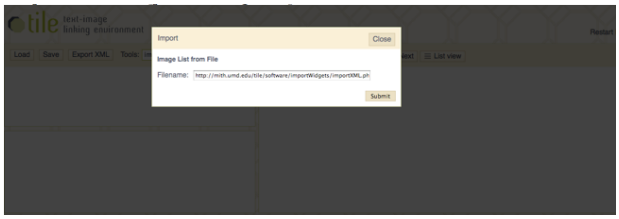


FIGURE A: Opening screen of TILE tool

TILE's JSON is simple and thoroughly documented, and we provide several translators to map common existing metadata formats to the format. We have already created a PHP script that will generate TILE JSON from a TEI P5 document and are currently working to do the same for the METS files used in the Indiana University's METS navigator tool. We hope to have other translators to show in our demonstration, particularly one that will import TEI/XML conformant with the University of Victoria's Image Markup Tool.

Additionally, TILE provides a modular exporting tool that allows users to run the work they've done in TILE through an external translator and then download the

result to the client computer. For example, a user may import a set of images and transcripts from a METS file at the Library of Congress, use TILE to link images and text, and then export the result as a TEI file. The TEI file may then be reimported to TILE at a later data to further edit or convert the file.

At the poster session at DH2011, we will demonstrate the final functionality of TILE and to display a poster and provide handouts that describe the thinking behind TILE, how it is intended to be used, and details on how TILE is built and functions.

References

Chymstry of Isaac Newton. http://www.chymistry.org.

Homer Multitext Project. http://www.homermultitext.org/.

Image Markup Tool. http://tapor.uvic.ca/~mholmes/image_markup/.

Mapas Project. http://mapas.uoregon.edu/.

METS Navigator. http://metsnavigator.sourceforge.net/.

Open Annotation Collaboration. http://www.openannotation.org/.

Swinburne Project. http://swinburnearchive.indiana.edu/.

Text Image Linking Environment. http://tileproject.org/.

CLAROS—Collaborating on Delivering the Future of the Past

Rahtz, Sebastian
sebastian.rahtz@oucs.ox.ac.uk
University of Oxford Computing Services

Dutton, Alexander
alexander.dutton@oucs.ox.ac.uk
University of Oxford Computing Services

Kurtz, Donna
donna.kurtz@beazley.ox.ac.uk
Beazley Archive, University of Oxford

Klyne, Graham
graham.klyne@zoology.ox.ac.uk
Department of Zoology, University of Oxford

Zisserman, Andrew
andrew.zisserman@eng.ox.ac.uk
Department of Engineering Science, University of Oxford

Arandjelović, Relja
relja.arandjelovic@chch.ox.ac.uk
Department of Engineering Science, University of Oxford

CLAROS (Classical Art Research Online Services) is a technology and data collaboration between classical art and archaeology research projects, museums and semantic web researchers. Documenting objects from the museums of the world, CLAROS aims to engage with the public across the widest possible spectrum. It builds on the success of the Beazley Archive which has provided programmes for the public as well as the scholar and an illustrated linked dictionary for more than fifteen years. CLAROS (http://www.clarosnet.org/) is based at the University of Oxford, with partners in the UK, France, Germany, Switzerland and Greece. The portfolio includes an aggregating RDF database, web discovery interfaces for different types of audience, visual search using image analysis of shapes and images, semantic information extraction from digitized text, place and name gazetteers, and investigation of avatars for resource discovery. CLAROS aims to be an effective and powerful partner in the realm of semantic web and linked data about the past.

1. Data Modelling

The CLAROS aggregating data cache pulls information from its partners, limiting itself to those interchange components which can be mapped to the CIDOC Conceptual Reference Model (CRM) ontology[1], with a few extensions relating to dates and geolocations. The majority of records so far use CRM concepts *E22_Man-Made_Object*, *E53_Place*, *E52_Time-Span* and *E21_Person*. The data contributions, using XML RDF as the ingest form, include the Beazley Archive (pottery and gems)[2] and the Lexicon of Greek Personal Names (onomastic data)[3] at the University of Oxford, the *Arachne* Sculpture Archive[4], the *Lexicon Iconographicum Mythologiae Classicae* in Paris[5] and Basel[6], and the German Archaeological Institute[7], producing an RDF triple store of over 20 million assertions. The federating and subsidiarity principle of CLAROS is that it acts simply as a resource discovery system, with search results linking back to the host database for more information, preserving the IPR and intellectual integrity of each partner. A SPARQL interface and RESTful APIs are provided for expert use, but CLAROS itself provides an exemplar query and visualisation interface (the Explorer) with an emphasis on textual search, timeline display and mapping of results. It is expected that the user will start with broad search terms, receive back information from a wide variety of sources, and then gradually refine and explore the results, perhaps ending in unexpected places.

2. Metamorphoses

One of the research subprojects of CLAROS is *Metamorphoses*, whose aim is to establish a working co-reference system for name, place and date information in classical art and archaeology and ancient history projects at Oxford. This will provide geo-naming and geo-locating services to both CLAROS partners in their normal research, and to the CLAROS Explorer in performing searches. One output is an aggregating Ancient Place Server with web services to answer queries about locations of places, names of places, and types of places within a chosen area. We work on a social model of places as objects which come into existence as a result of naming by one or more groups of people at a particular time. We expect places to have multiple names, to have different geographical limits over time, and to have relationships with other places.

3. Visual Search

A novel aspect of CLAROS is that visual queries can be used to access and search the archives that are linked to within CLAROS. Suppose, for example, that a novice takes a photograph of a Greek vase in a museum on their iPhone, or finds an image of a Greek vase on the web, that they would like to know more about. This image can be use as a visual query to retrieve that vase from the archive, in much the same way as the text phrase “Greek vase” can be used as a query in Google to retrieve web pages which contain that phrase. The method is illustrated in Figure 1: the image is uploaded to a web server, and the server returns the matches in the archive together with meta-information – for instance identifying the type of the vase, its date, its material, its decorations etc.

Shape: amphora, neck;
Fabric: athenian;
Technique: black-figure;
Date: 550-500;
Artist: mastos group;
lysippides p;
Scholar: beazley;
beazley;
Decoration: dionysos with kantharos and grapevine, between hermes and satyr;
ajax with body of achilles, warrior, boeotian shield;

Figure 1: Identifying a vase from a web photo: an image of a vase is uploaded from a web link (top), and is matched to one in the Beazley vase collection (middle) in real time (a search through over 100,000 images). This identifies the type of vase, its material, date and decorations from the meta-information (bottom) associated with this data in the Beazley archive.

Figure 2 gives another example, this time for a visual search of the *Arachne* classical sculpture archive.

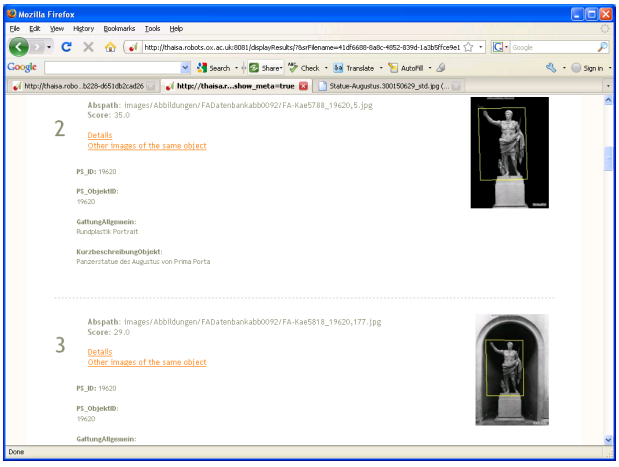


Figure 2: Identifying a sculpture from a web photo: an image of a sculpture is uploaded from a web link (top), and is matched to one in the Arachne sculpture collection. In turn, this provides links to other images of the sculpture in the collection and also to associated meta-information (bottom).

That this visual search is possible, and indeed can be carried out with results being returned immediately, is due to recent methods developed in the computer vision community on visually searching for objects in large scale image datasets (see [1] for details).

References

J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman (2007). 'Object Retrieval with Large Vocabularies and Fast Spatial Matching'. *IEEE Conference on Computer Vision and Pattern Recognition*.

Notes

1. http://www.cidoc-crm.org/
2. http://www.beazley.ox.ac.uk/
3. http://www.lgpn.ox.ac.uk/
4. http://www.arachne.uni-koeln.de/drupal/
5. http://www.mae.u-paris10.fr/limc-france/
6. http://www.limcnet.org/Home/tabid/77/Default.aspx
7. http://www.dainst.org/

Interactive Layout Analysis, Content Extraction and Transcription of Historical Printed Books using Agora and Retro

Ramel, Jean-Yves
ramel@univ-tours.fr
Lab d'Informatique, Ecole Polytechnique de l'Université de Tours

Sidère, Nicholas
nicolas.sidere@univ-tours.fr
Lab. d'Informatique, Ecole Polytechnique de l'Université de Tours

High level analyses of document images are mainly based on the output of a page segmentation process. For example, the extracted text regions can be the input to an OCR system to retrieve the ASCII characters printed on the pages. The spatial relationships between segmented blocks along with other features can be used in logical page organization analysis to group the extracted components appropriately and recover the correct reading order. Many techniques for page segmentation have been proposed in the literature but most of them are based on the assumption that an input document image consists of a set of rectangular blocks. Furthermore, the classification step is generally domain specific and uses static rules to automatically determine, for each block, the coherent label selected from a predefined list (title, paragraph, graphic, table,...). These limitations appear too restrictive with respect to some noisy and distorted documents and new approaches need to be developed.

In this context, we present a work achieved in collaboration with the “Centre d’Etude Superieur de la Renaissance” of Tours (CESR / http://www.cesr.univ-tours.fr). The CESR is a training and research centre which receives students and researchers who wish to work on various domains of the Renaissance using a rich library of historical books. The CESR wants to create a Humanistic Virtual Library; however, until now, only bitmap versions of several books that have been scanned or photographed are accessible. The initial objective of the CESR was to obtain an ASCII version of the text contained in the pages of these historical books. The centre first tried to use the commercial OCR software to index their books but they

quickly realized that, applied to historical documents, this procedure would had been vowed to failure. So, the CESR asked our Pattern Recognition and Image Analysis research team to help them to define a new system adapted to their needs. They have appreciated our efforts as our collaboration will lead to a system able to bring a better description and indexation of the content of their books and would also make the search and the reading of these precious historical books easier.

The poster will first describe the new hybrid method we have developed for the extraction of layout information and of specific elements like graphical parts or ornaments based on the construction of two representations of the contents of the images. A mapping of the shapes and a mapping of the background are computed. By exploiting this information, our algorithm produces and sends back a list of blocks constituting a first segmentation result. Then, this initial representation of the image is used during a more sophisticated analysis. Having an aim of genericity, the architecture of the system that we carried out authorizes an interactive installation of scenarios for analysis of the image contents. Scenarios work on the initial representation provided by the first step of the segmentation. According to its needs (localization of the ornamental letters, the notes at margins, titles,…) and using user-friendly interfaces, the user (not expert in image processing) builds scenarios allowing to label, to merge, to remove the blocks contained in the intermediate representation. One can thus locates the desired entities without taking care of the other areas of the image. The elaborated scenarios can then be stored, modified and applied to various sets of images during batch processing. The results obtained with this method are very interesting; the adjustment of the necessary parameters is straightfoward and not sensitive to variations. The originality of our approach lies in the opportunity which we offer to the users to be able to build, in an interactive way, scenarios of incremental analysis. We propose to call this new method “user-driven analysis” in opposition to data-driven or model-driven methods. The goal is, on the basis of the initial segmentation, to be able to make the representation of the images evolve in a progressive way to lead to the finest possible characterization of its contents according to the user objectives and to the type of images to be analyzed. The CESR has processed several complete books using AGORA prototype and their own scenarios of block classification. Thus, the CESR has increased the number of books offered to the users in its Virtual Library (see http://www.bvh.univ-tours.fr). Even if the system produced some errors, the processing and time saved as compared to manual processing is considerable (for exemple, the manual indexation of the page layout of an historical book of 300 pages last approximately two days instead of only two hours when using Agora), this providing to the specialists of historical books, a useful tool which they had never imagine (see Figure 1).

Concerning text transcription, the originality of our work relies upon the analysis and exploitation of pattern redundancy in documents to allow efficient and quick transcription of books as well as identification of typographic materials. This pattern redundancy is mainly obtained via clustering methods. Like this, the traditional OCR problem could be reformulated into a text transcription one. A text, be it ancient or not, is made up of sequences of symbols. The scanning process produces pictures where symbols are represented as thumbnails of patterns (a pattern could be a single character, a part of a character or a set of joined characters), which may be more or less distinct. Without prior knowledge about the meaning of these symbols, the application of a clustering approach assigns thumbnails of a similar shape to the same cluster. As an example, one cluster containing thumbnails of the lowercase letter “a,” another one the uppercase letter “A,” yet another one the letter “b” in a specific font, and so on. Once the clustering is done, a user could assign a label to each cluster using a other Graphics User Interface (software called RETRO). These labels are then automatically assigned to each pattern, thus achieving the text transcription of the whole book. In this way, if 90% of patterns are detected as redundant, only one character in ten will be labeled by the user in order to transcribe the book. This part of the work is still in progress and is corresponding to a Google Digital Award obtained by our team in December 2010.

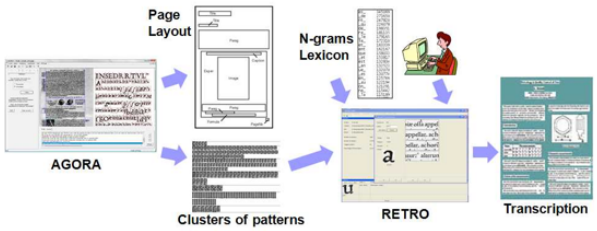


Figure 1 : A view of the proposed processing framework with Agora and Retro

Enhancing Museum Narratives: Tales of Things and UCL's Grant Museum

Ross, Claire
claire.ross@ucl.ac.uk
University College London

Hudson Smith, A.
a.hudson-smith@ucl.ac.uk
University College London

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

Warwick, Claire
c.warwick@ucl.ac.uk
University College London

Carnall, Mark
mark.carnall@ucl.ac.uk
University College London

Emergent mobile technologies offer museum professionals new ways of engaging visitors with their collections. Museums are powerful learning environments and mobile technology can enable visitors to experience the narratives in museum objects and galleries and integrate them with their own personal reflections and interpretations. UCL's QRator project is exploring how handheld mobile devices and interactive digital labels can create new models for public engagement, personal meaning making and the construction of narrative opportunities inside museum spaces.

The QRator project is located within the emerging technical and cultural phenomenon known as 'The Internet of Things': the technical and cultural shift that is anticipated as society moves to a ubiquitous form of computing in which every device is 'on', and connected in some way to the Internet. The project is based around technology developed at the Centre for Advanced Spatial Analysis, UCL and is an extension of the 'Tales of Things' project which has developed a 'method for cataloguing physical objects online which could make museums and galleries a more interactive experience' (Giles, 2010) via means of QR tags.

The project aims to genuinely empower members of the public within the Grant Museum by allowing them to become the 'Curators'. The project develops a custom UCL Museums iPhone, iPad and Android application which will be available free of charge from the iTunes store and Android market place.Small printed QR codes for museum objects will be created, linked to an online database allowing the public to view 'curated' information and most notably to send back their own interpretation and views via their own mobile phone. Unique in the UCL technology is the ability to 'write' back to the QR codes. This allows member of the pubic to type in their thoughts and interpretation of the object and click 'send'. Similar in nature to sending a text message, the system will enable the Grant Museum to become a true forum for academic-public debate, using low cost, readily available technology, enabling the public to collaborate and discuss object interpretation with museum curators and academic researchers. Visitors narratives subsequently become part of the museum objects history and ultimately the display itself via the interactive label system to allow the display of comments and information directly next to the artifacts.

QRator provides the opportunity to move the discussion of objects from the museum label onto users' mobile phones, allowing the creation of a sustainable, world leading model for two-way public interaction in museum spaces. UCL's Grant Museum of Zoology houses one of the country's oldest and most important natural history collections. The Grant museum has a strong history as a teaching collection but also functions as a key gateway for the public to engage with academic issues in innovative ways.

Museums have undergone a fundamental shift from being primarily a presenter of objects to being a site for experiences, which offer visitors opportunities for individual meaning making and narrative creation. Many visitors expect or want to engage with a subject, physically as well as personally (Adams et al 2004; Falk and Dierking 2000). Visitors see interactive technology as an important stimulus for learning and engagement (Falk et al 2002; Black 2005), empowering users to construct their own narratives in response to museum exhibits. Beyond expected content synthesis, these immersive activities can stimulate learning. Engaged within this immersive environment, museum objects become rich sources of innovation and personal growth (Fisher and Twiss-Garrity 2007). When visitors experience a museum which actively encourages individual narrative construction, their activity is directed not towards the acquisition or receipt of the information being communicated by the museum, but rather towards the construction of a very personal

interpretation of museum objects and collections. The unpredictability of multiple narrative forms created by the use of mobile devices and interactive labels introduces new considerations to the process by which museums convey object and collection interpretation and opens up museums to become a more engaging experience.

The participation in collaborative narrative creation centred around museum objects can provoke creative, independent analysis, promoting a personal connection with museum exhibition subject matter that is unparalleled in more traditional and passive approaches (Silverman 1995; Roberts 1997; Hooper-Greenhill 2000; Fisher and Twiss-Garrity 2007). This research aims to stress the necessity in actively engaging visitors in the creation of their own interpretations of museum collections. This poster presents the development of the QRator project so far, highlights the user centred development activities, its opportunities, challenges and provides an insight into how utilising mobile technology can enhance visitor meaning making and narrative construction.

References

Black, G. (2005). *The Engaging museum: Developing museums for visitor involvement*.

Falk, J. H., Dierking, L. D. (2000). *Learning from the Museum: Visitor Experiences and Making Meaning*.

Falk, J. H., Cohen Jones, M., Dierking, L. D., Heimlich, J., Scott, C., Rennie, L. (2002). 'A multi-institutional study of exhibition interactives in science centers and museums'. *Unpublished evaluation report. Annapolis, MD: Institute for Learning Innovation*.

Fisher, M., Twiss-Garrity., B.A. (2007). 'Remixing Exhibits: Constructing Participatory Narratives With On-Line Tools To Augment Museum Experiences'. *Museums and the Web 2007: Proceedings. Toronto: Archives and Museum Informatics.* `http://www.archimuse.com/mw2007/papers/fisher/fisher.html`.

Giles, J. (17th April 2010). 'Barcodes help objects tell their stories'. *New Scientist*.

Hooper-Greenhill, E. (2000). 'Museums and the interpretation of visual culture'. *Museum Meanings Series*.

Roberts, L.C. (1997). *From Knowledge to Narrative: Educators and the Changing Museum.*.

Silverman, L. H. (1995). 'Visitor meaning making in museums for a new age'. *Curator.* 38(3): 161-169.

Documenting Horizons of Interpretation in Philosophy

Saisó, Ernesto Priani
epriani@gmail.com
History of Philosophy and Humanities Computing, Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México

Farfán, Leticia Flores
floresfarfanleticia@gmail.com
History of Philosophy, Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México

Zavala, Daniel
siedrix@gmail.com
Independent software developer

Choreño, Rafael Gómez
rafaelangelg@yahoo.com
History of Philosophy, Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México

Priego, Ernesto
efpriego@gmail.com
Department of Information Studies, University College London

1. Abstract

URL of the project:`http://siedrix.com/work/acms/`

This poster presentation will describe a project currently underway at the Faculty of Filosofía y Letras, Universidad Nacional Autónoma de México (UNAM). It has been funded by a grant from the Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza (PAPIME), a university endowment program for innovation in education. It is a collaborative documentation project of the individual and collective processes of research and writing of a group of 25 researchers, working on papers about the contemporary strategies of appropriation of the antiquity and aims to use innovative web-based technology to document different data of the research/ writing process and to create a visualization model to use the data for educative and didactical purposes.

2. Background, Assumptions and Needs

In May 2010 a group of professors and researchers teaching at the Facultad de Filosofía y Letras, UNAM, gathered to start a project following Barbara Cassin's proposal from the conference *Les Stratégies contemporaines d'appropriation de l'Antiquité*. The objective is to create a seminar and produce papers on the subject of the contemporary strategies of appropriation of the antiquity, for a series of publications to be used as a teaching material in a variety of courses in Contemporary and Classical philosophy at the faculty.

The main assumption of the project is that we cannot have access to the classic antiquity without mediation or heritage (Gadamer 2005). Reading and studying Classical philosophy implies the act of taking a position in a horizon of interpretation from which the object of research is built, the problems around it are articulated and the center of attention is decided.

Project's seminar and papers will discuss those strategies in contemporary authors from France and Germany, who study the antiquity and point out the horizon of interpretation from which they approach the ancient texts. The findings will be used to teach the conceptions of contemporary authors and their ideas of Classical philosophy, as an example of Nietzsche's thesis that there are not facts but interpretations (1998), and to explain technically to students the process that produces an interpretation.

The idea to use a web-based technology for a collaborative documentation project came afterwards, once all these ideas about the project were settled. A paradox emerged from the project approach: in order to describe those horizons of interpretation of all the researchers, both as individuals and as a group, needed to create their own horizon of interpretation. Can we document that process? With that documentation, can we answer the question "from which horizon are they writing?" And can we use the documentation for teaching?

As opposed to traditional research in philosophy, which analyzes finished works to set the horizon of interpretation; we want to produce a record of an ongoing work to show how the horizon is created and how it changes with the influence of new readings or the interaction with other researchers. These goals lead us to two main problems:

1. What kind of data do we need to collect to have an idea of the horizon of interpretation?
2. How to collect data in an open and dynamic way that allows us to follow the process of research and its changes during a period of time?

We focus on the two main activities in academic literacy: reading and writing. The kinds of data we want

to obtain are the reading annotations as an evidence of reading and the variants of the progressive writing[1]. Also, it will be relevant for us to keep a record of the methodology adopted by the individual researchers as a single data or as a succession of variants.

All these data must be obtained while researchers are working on the project and until they finish their papers. It is vital to have not only a record of individual activity but also of the collective work of the professors.

3. Blogging, Annotating, Writing

Our project is an application of collaborative "crowdsourcing" (Albors 2008), following previous experiences in Digital Humanities. For instance, we have in mind the success of “A Day in the Life of Digital Humanities”, using blogging and wiki for collaborative documentation of daily activities of digital humanists[2]. It consists of each one of our 25 researchers writing their own blog in which they document their daily research activities, as well as commentating on other blog entries during a year. This mechanism will allow us to have the seminar virtually and to trace collective work around independent participation. However, additionally we need to collect annotations and methodological assumptions. In order to do so we have already developed a web application for these specifics needs:

3.1. Capturing and Highlighting Quotes:

With our web application researchers can upload book passages and highlight some paragraphs in it, to document and share its readings. Every passage will be able to be marked by different researchers and the web application allows comparisons between these marks. Highlighting different parts of a same text shows individual decision and collective assumptions.

During the process of the project, all the mark up will be made as a reference to a place on the passage and they will be part of a rendering system in Javascript. This makes the marks live outside the text. Once the project is finished, the marks will be exported to TEI to preserve the documents and marks that will be part of the text. Nevertheless, during this process each contributor can mark Dates, Names, Terms and Quotes as done with TEI Lite in order to also produce a strong markup of the text to use it didactically for students, for example, by linking those marks to complementary information about the persons named, the terms used, etc. Passages entries and highlights can be commented by other participants and can be linked to blog entries.

3.2. Explain and Update Methodology

Researchers can explain their methodology and are able to update their methodological assumptions anytime they decide to change the study methodology or the study subject in our web application, simply by filling the methodology box in their profile. At the end of the project we will be able to compare the evolution of the research in a timeline, mixed with their blog post and annotation, having an insightful way to analyze the initial work and their finished paper.

4. Individual and Collective

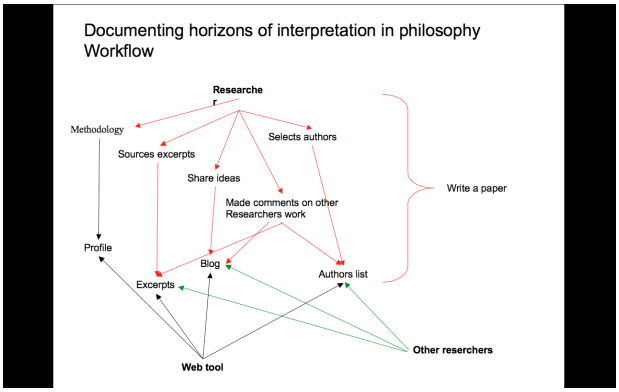
Highlighting quotations and comments on blog and quotation entries are ways to have evidence of individual decisions and collective assumptions. We extend this view to the creation of metadata. As it is used in many commercial and academic projects, researchers can create metadata or use those that had been created by others, for sources excerpts and blogs. Metadata are central to finding links between autonomous and collective work. The final ontology can help us to understand how strong the collective or the individual view has been. We are still working on the metadata creation rules.

5. The Didactic Aim

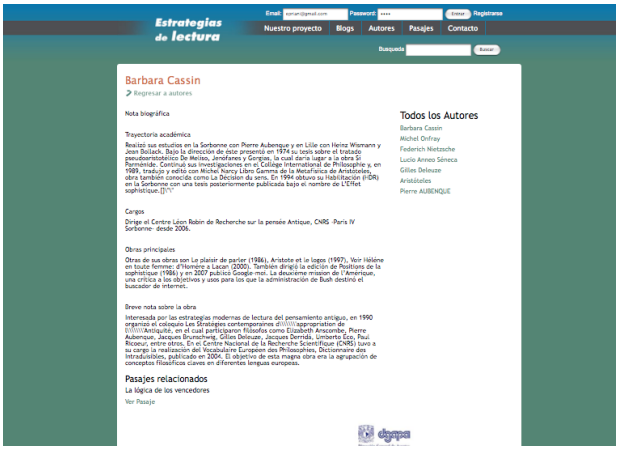
As an educational project, the last objective is to use the documentation to help students to understand how we study the classical antiquity from a horizon of interpretation. As a tool for education, it allows students to:

1. Follow individual researcher’s work in real time during the activity of the project.
2. Compare quotes and highlights being used by different researchers.
3. Follow the entries (blogging, passages, highlighting, methodology) on a modern or classical author.
4. Have extra information about modern or classical authors (reference, biography, bibliography)

6. Workflow

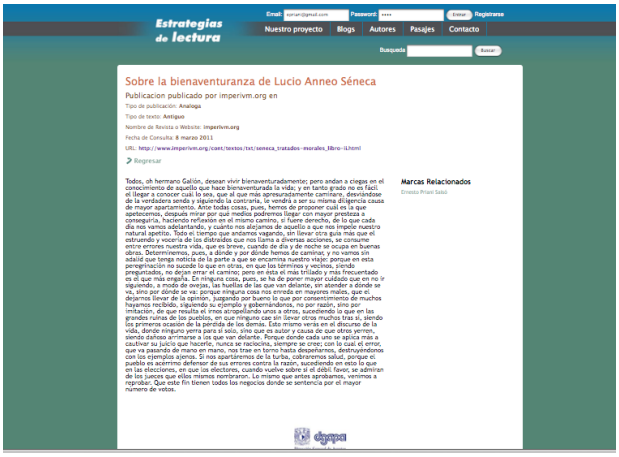


7. Authors



8. Blog

9. Sources Excerpts



References

Albors, José et al. (2008). 'New learning network paradigms: Communities of objectives, crowdsourcing, wikis and open source'. *International Journal of Information Management.* 3: 194-202.

Burnard, Lou , C. M. Sperberg-McQueen (2006). *TEI Lite: Encoding for Interchange: an introduction to the TEI — Revised for TEI P5 release.* `http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html`.

Gadamer, Hans-Georg (2005). *Verdad y Método I [Traducción de A. Agud y R. de Agapito].* Salamanca: Sígueme.

Howe, Jeff (2006). 'The Rise of Crowdsourcing'. *Wired.* June 2006. `http://InformationR.net/ir/14-3/paper410.html`.

Nietzsche, Friedrich (1998). *Sobre verdad y mentira en sentido extramoral, seguido de La voluntad de ilusión de Nietzsche de Hans Vaihinger [Traducción de Luis M. Valdés y Teresa Orduña].* Barcelona: Tecnos.

Notes

1. Cheryl Geisler (1994). Academic Literacy and he Nature of Expertise: Reading, Writing, and Knowing in Academic Philosophy, Laurence Erlbaum Associates, Publishers. New Jersey, 1994 pp. 15 ss
2. Day in the Life of the Digital Humanities project is at http://tapor.ualberta.ca/taporwiki/index.php/Day_in_the_Life_of_the_Digital_Humanities_2010

Visualization of Visitor Circulation in Arts and Cultural Exhibition

Sookhanaphibarn, Kingkarn
kingkarn@ice.ci.ritsumei.ac.jp
Intelligent Computer Entertainment Laboratory,
Global COE Program in Digital Humanities for
Japanese Arts and Cultures, Ritsumeikan University

Thawonmas, Ruck
ruck@ci.ritsumei.ac.jp
Intelligent Computer Entertainment Laboratory,
Global COE Program in Digital Humanities for
Japanese Arts and Cultures, Ritsumeikan University

Rinaldo, Frank
rinaldo@is.ritsumei.ac.jp
Intelligent Computer Entertainment Laboratory,
Global COE Program in Digital Humanities for
Japanese Arts and Cultures, Ritsumeikan University

1. Introduction

The topic of visitor circulation in museums and art galleries has been considered as an important factor in all aspects of the museum experience (Bigood, 2006; Guy *et al.* 2010; Kaynar *et al.* 2009) Circulation describes how visitors explore a set of exhibits in a particular space by observing what pathways the visitors take. A visualization of visitor circulation can confirm whether visitors circulate the way the designers intended. The visualization can assist the designers to arrange a predefined pathway so that visitors will not miss key exhibits. The well-designed circulation system can also increase the great number of return visitors.

Sookhanaphibarn and Thawonmas (2009) proposed the local and global visualizations aims at presentation and analysis of visitor behaviors in 3D virtual museums. The visitor path is displayed with a spectrum of colors in the form of connecting segments from red to violet. The color of a particular segment indicates the passage of time. The drawback of these visualizations is that they were strictly overlaid on the layout map. To deal with the varying layouts commonly found in a museum with many exhibition rooms, visualizations with an independent layout is an alternative assistant tool for the visual analytics of circulation patterns.

In this paper, we proposed a new visualization tool to represent a visitor path and his/her time spent residing near the closest item. We encode a time interval residing in an item boundary into a color-shaded line segment. Color shade is used as an indicator of the proximity to the nearest item. The length of a segment is in proportion to the total time spent in the layout. The time segment is placed in the row corresponding to its item boundary. A path of visited items is illustrated by connecting the time segments with vertical lines.

With the proposed visualization, we can easily find the trend of visitor circulation which strictly follows the pathway designed by a curator. The trend is represented by the white line, called “Forward”, running from the most left above to the most right bottom corners as shown in Figure 1. There are the other trends, which are called “Backward”, “Bell” and “Inverted bell”, possibly influenced by the visitor characteristics and/or preferences. The similar trends of visitor circulation will be explained later in detail.

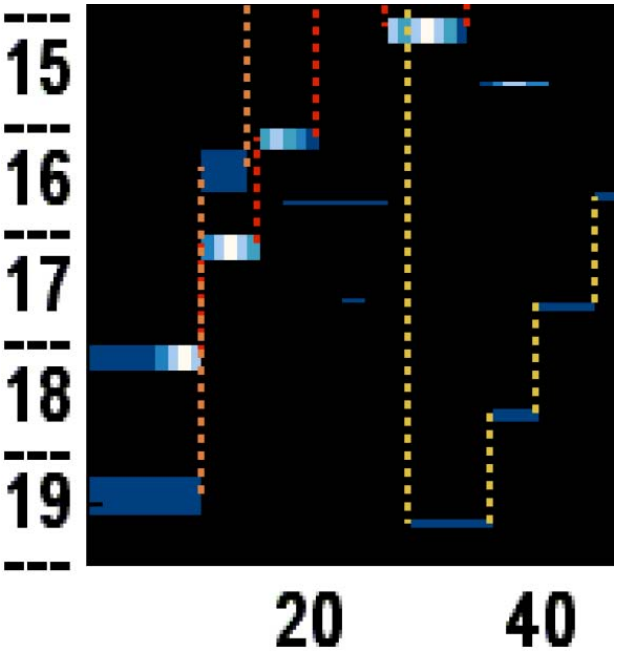


Figure 1: Four trends of circulation: (a) Forward, (b) Backward, (c) Bell, and (d) Inverted bell.

1.1. Layout-dependent visualization

Most visualization techniques using traditional two dimensional maps represent visitor trajectories and their corresponding visit time directly over spatial layouts. In this paper, we call layout-dependent visualization as any visualization technique using the spatial layout of a target area as its graphical background in visualization. With the layout-dependent visualization, a circulation pattern is not easily extracted by a user who is not familiar with the target layout map.

The layout-dependent visualization approach burdens users with a cost of requiring them to recognize the layout of items and routes by themselves. Some additional symbols indicating the position and boundary of items as well as arrowheads indicating representative routes must be placed in layout-dependent visualization. However, these symbols and arrowheads conceal visitor traces from users.

2. Design Decision

This paper considers the use of a layout-independent display for visual analysis of the path and residing time of the movement data in circulation, named “PARTY”. PARTY is an abbreviation of Path And Residing Time display. Taking for example a museum in a 3D virtual world, the circulation behaviors of visitors moving through museum of interest are influenced by the items on which the visitors focus their attention.

Designers of a musuem space require serveral types of information when examining the circulation behavior of visitors. These include (1) residing locations, regions, or item boundaries, (2) visit time intervals near an item, (3) paths of visited regions, (4) global information showing multiple visitors residing in a region, and (5) degrees of their interest. All of these information types are derived from two data sources: (a) a log file of visitors’ positions including x-y coordinates and time and (b) the map of a museum or a floor plan, where the location of items or the position of rooms is provided at least.

The design of PARTY aims to represent three dimensional entities, i.e. a time unit, a visited item, and a visitor. The horizontal and vertical axes of PARTY represent time and items which visitor moves through their boundaries. A stack of visitor stripes is put inside an item block (row). As shown in Figure 2, there are five items of interest and three visitors. The stack of three visitor stripes places inside each item row. The order of three visitor stripes is consistent thorough five item rows. Therefore, an PARTY entity represents a visitor v who stays near item X at time t.

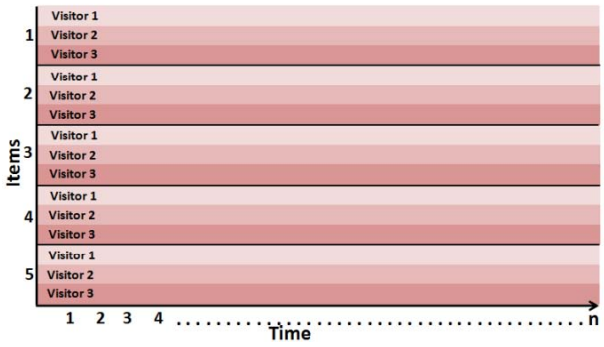


Figure 2: A structural design of PARTY.

We arrange the visitor stripes in every stack in the same order and rank them by the similarity among their present circulation patterns. Then, we use the hierarchical clustering technique for finding the similarity of all pairs of visitor paths. To handle hundreds of visitors, the representative of each group of visitors can be used rather than a single visitor. The representative is derived from the generalized median defined as the visitor path having the nearest distances to all.

Besides a path of nearest items versus time, the proximity distance to the nearest item can be displayed in PARTY as a degree of visitor’s interest to each exhibit item. A displaying color is computed using the observation distance and range based on the location of all items. For example, given the visitor trace as shown in Figure 3, Figure 4 represents a visitor trace as our observation-based time series. The visualization in Figure 4 consists of the horizontal axis corresponding to the visit time, the vertical axis indicating an item belonging to the observation range at a particular time, and the color showing the observation distance. This visualization is produced by using the observation distance and range in the application of a 3D virtual museum.

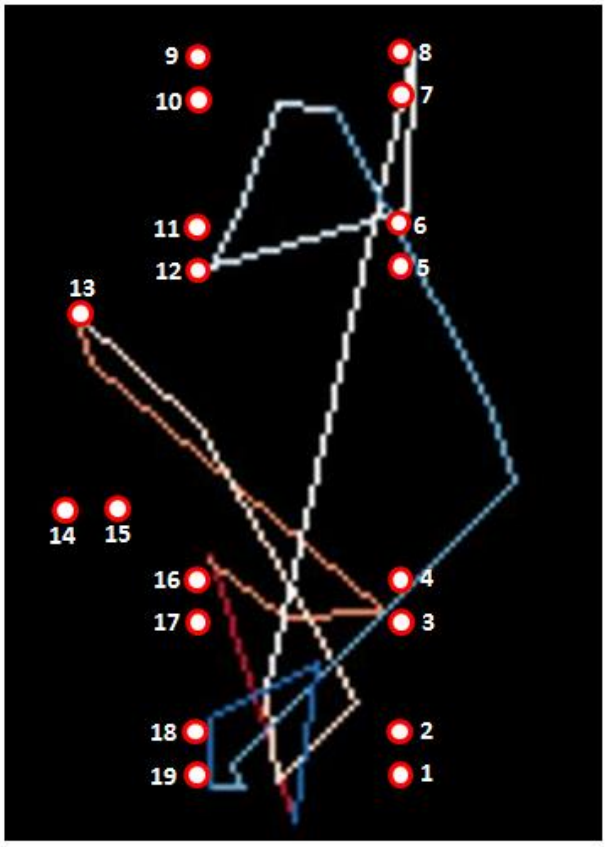


Figure 3: A representation of a visitor trace tracked in a 3D virtual museum.

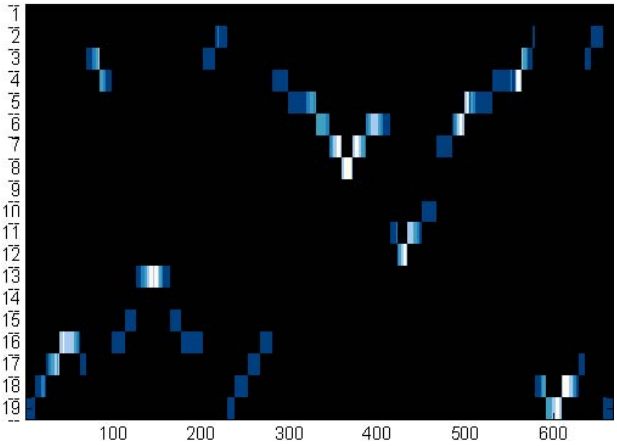


Figure 4: Transformation of a single trace (Fig 3.) to PARTY representing the path of visited items (y-axis) and his/her interest to items by using color shades from brightness to darkness (highest to lowest degrees).

3. Results and Implications

This section presents an application of PARTY analyzing the avatars' trajectories and finding trends of circulation behaviors in the 3D virtual museum, named RDAP. RDAP, owned by the Global Center of Excellenct in Digital Humanities Center for Japanese Arts and Cultures, of Ritsumeikan University, was created in Second Life. An objective of RDAP is to disseminate Japanese costumes, Kimonos, preserved them in a digital achieving system. We synthesize the visitor trajectories based on the metaphor of four animals as mentioned in (Sookhanaphibarn and Thawonmas, 2009). The total number of synthesized trajectories is 36 where each visiting style has nine trajectories.

3.1. Similar Trends of Circulation Patterns

After applying the PARTY approach, visitor trajectories are transformed into time series data. Discovering similar trends of circulation patterns is achieved by a traditional dynamic time warping followed by the hierarchical clustering. Then, generalized median of each resulting cluster is calculated. Figure 1 shows four trends of circulation patterns including:

- a) Forward circulation: this trend illustrates that visitors prefer to turn right at entrance and move following the curators-guided path from the the fist to the last items.
- b) Backward circulation: the direction of this circulation is backward from the forward one, i.e. turn left at the beginning.
- c) Bell-shape circulation: visitors in this trend prefer to start and end their visit with the same item, and they turn right at entrance.
- d) Inverted bell-shape circulation: This trend pattern is similar to the bell circulation but they turn left at entrance.

The other view of PARTY displays a stack of representative stripes of which the width denotes the size of their categories. The bell-shape circulation has the largest number of visitors followed by the backward, forward, and inverted bell-shape circulations, respectively.

References

Bitgood, S (2006). 'An Analysis of Visitor Circulation: Movement Patterns and the General Value Principle'. *Curator: The Museum Journal.* 49: 463–475. `http://doi:10.1111/j.2151-6952.2006.tb00237.x`.

Guy, G., Dunn, S., Gold, N. (2010). 'Capturing Visitor Experiences for Study and Preservation'. *Digital Humanities 2010, Conference Abstracts Book.* V. 2, pp. 160-16.

Kaynar, I.R., Psarra, S., Wineman J. (2009). 'Experiencing museum gallery layouts through local and global visibility properties in morphology: an inquiry on the YCBA, the MoMA and the HMA'. *Proc. 7th International Space Syntax Symposium.* Stockholm, Sweden, June 2009, pp. pp. 1-16.

Sookhanaphibarn, K., Thawonmas, R. (2010). 'Visualization and Analysis of Visiting Styles in 3D Virtual Museums'. *Digital Humanities 2010, Conference Abstracts Book.* Pp. 239-243.

Mashing up the Map: Film Geography and Digital Cartography in a Cultural Atlas of Australia

Stadler, Jane
j.stadler@uq.edu.au
University of Queensland

This poster presents research from the online movie map component of a new digital resource: A Cultural Atlas of Australia is a collaborative, interdisciplinary digital humanities research project that uses interactive cartography and spatial theory to map Australian narrative fiction across three media forms. Building on the growing interest in digital cartography and spatial theory in the humanities, this initiative investigates the cultural and historical significance of location and landscape by presenting a national survey of narrative space spanning Australian novels, films and plays. The broader project involves Dr Peta Mitchell from literary geography and Dr Stephen Carleton from theatre studies in an investigation of the mediation, remediation and geovisualisation of locations and landscapes in cinematic, literary, and theatrical narratives.

There has been a recent surge of popular and critical interest in linking online mapping with cinema. This manifests most obviously in a rash of “movie maps” and online spatial resources such as Robert Allen’s Going to the Show, which digitally maps a history of moviegoing in North Carolina (`http://www.docsouth.unc.edu/gtts`)[1]; the National Film and Sound Archive’s film-location map on the Australian Screen Online site (`http://aso.gov.au/`); and Sébastien Caquard’s work on cinematic cartography in the Canadian Cinematographic Territories Atlas (`http://www.atlascine.org`), which traces the history of animated maps and virtual globes such as Google Earth through cinema, examines the technological interface connecting cinema and cartography, and maps the locations of cinemas across Canada[2]. While many such projects focus on mapping film production and distribution locations, I seek to focus attention on developments in movie mapping from since 1980s and situate such work in relation to cinematic cartography and the complex interplay of narrative settings and shooting locations in Australian cinema.

Over the past decade, interactive online mapping —what D. R. Fraser Taylor has called "cybercartography"[3]—has become a particularly salient issue within cartography, geography, and humanities research.[4] Since Google released its Google Maps API in 2005, the Google Maps code has been freely accessible as long as the resulting map "mashup" remains nonproprietary and in the public domain. According to William Buckingham and Samuel Dennis, the development of open source mapping tools, such as Google Maps, has generated much interest in the use of maps for "understanding 'non-mapped' phenomena (e.g., qualitative data or localized community information and knowledge)."[5]

Recent technological developments in digital cartography make it possible to visualise and to map the ways in which spatial storytelling produces and translates space across different media. This poster presents geovisualisation—in the form of an interactive online map—as a means by which to map representations of iconic landscapes and sites within Australian cinema. The Cultural Atlas of Australia uses an interdisciplinary method incorporating cultural geography and textual analysis to interpret, situate, and contextualise the representation of location. The project remediates that information in the form of an online, interactive map that has the potential to suggest new ways of thinking about location and landscape and break down traditional typologies of Australian space. This digital humanities research advances on traditional cartographic explorations and representations of space, making it possible to visualise new perspectives on, intersections between, and layerings of geographic and textual information. This aims to enable the identification of regional tropes, patterns and gaps in spatial representations that may not have previously been evident in research that focused on isolated case studies of individual texts, whether literary, cinematic or theatrical.

As a research tool, the finished map will be searchable (by medium, location, theme, author, and text) and will enable users to generate and export their own maps with information they require. These functions have the capacity to reveal how cultural meanings accrue on the landscape, and how our relationship with, and understanding of, the natural and cultural environment changes over time. The possibilities for incorporating participants' photographs, videos, and textual accounts of Australian places via mobile social computing technologies opens up still more opportunities for the representation of multiple perspectives.

Geovisualisation has the potential to pose new questions for spatial analysis and to encourage broader public engagement in cultural geography. However, as a form of remediation, it does carry its own representational problems. As Barbara Piatti et al have noted, the geography of fiction is an imprecise one.[6] Piatti is speaking about mapping literary fiction, but the point can be made for all forms of narrative fiction. The representation of space and place can never simply be mimetic, but always, to a greater or lesser degree, creates an imaginative geography that may correspond to what Piatti calls the "geospace" (or map space)[7] directly, obiquely, or not at all. Bringing film space into the analytical frame carries its own set of complexities and ambiguities because film requires attention to the relationship between narrative locations and shooting locations. Beyond the question of impreciseness, the process of re-presenting narrative locations is a process of imagining and re-imagining geography that, by its very nature, is also political.

This project demonstrates that films are more than representations, more than containers for narrative symbolism and ideological views and values, and this extends to any geovisualisation strategy that seeks to map those texts. Such texts are also generative —productive of meanings, social relationships and subject positions. Tom Conley argues in *Cartographic Cinema* that cinematic images "produce space through the act of perception";[8] similarly, film stages and imaginatively invokes space in ways that subsequently inflect the meanings readers associate with actual places. Where film geography and cinematic cartography enable analysis of locational information in narrative fiction informed by insights from geography as well as cinema and cultural studies, it also builds from the premise that such texts intervene in the cultural field and alter the perceptual, ideological, political and practical orientation of readers and audiences in relation to the physical environment.

Notes

1. Allen, Robert C. "Getting to Going to the Show," New Review of Film and Television Studies 8, no. 3 (2010): 264–276.
2. See Sébastien Caquard, "Foreshadowing Contemporary Digital Cartography: A Historical Review of Cinematic Maps in Films," The Cartographic Journal 46, no. 1 (2009): 46–55.
3. D. R. Fraser Taylor, Cybercartography: Theory and Practice (Amsterdam: Elsevier, 2006)
4. See, for instance, William Cartwright, Michael P. Peterson, and Georg Gartner, eds, Multimedia Cartography (Berlin: Springer-Verlag, 1999); Mark Monmonier, "Cartography: The Multidisciplinary Pluralism of Cartographic Art, Geospatial

Technology, and Empirical Scholarship," Progress in Human Geography 31, no. 3 (2007): 371–79; and Jeremy W. Crampton, "Maps 2.0: Map Mashups and the New Spatial Media," in Mapping: A Critical Introduction to Cartography and GIS (Malden, MA: Blackwell, 2010), 25–38.

5. William R. Buckingham and Samuel F. Dennis, Jr., "Cartographies of Participation: How the Changing Nature of Cartography has Opened Community and Cartographer Collaboration," Cartographic Perspectives 64 (2009): 55.
6. Barbara Piatti, et al., "Mapping Literature: Towards a Geography of Fiction," in Cartography and Art, ed. William Cartwright, Georg Gartner, and Antje Lehn (Berlin: Springer, 2009): 182.
7. Ibid.
8. Tom Conley, Cartographic Cinema (Minneapolis: University of Minnesota Press, 2007), 20.

Better Software Tools for the Humanities and the Social Sciences: a Computer Science Perspective

Stephenson, Russell
steprus2@isu.edu
Idaho State University

Kantabutra, Vitit
vkantabu@computer.org
Idaho State University

As computer use has become more prevalent in all areas of academia, more and more scholars in the humanities and the social sciences have begun to realize the potential usefulness of software tools for their respective areas of scholarship. Many have become expert in the use of software packages ranging from database management systems and GIS to graphical and Web-authoring programs, not to mention the more common packages for word processing and the creation of presentations.

These scholars have also, however, discovered the limitations of current computer software. The problems range from lack of user friendliness to gross inefficiency and the inflexibility of database systems, to the point that important analytical discoveries which should be made are not made. We believe that these problems occur because these software packages were created to serve limited business purposes, and were co-opted for academic use simply because there were no better alternatives available.

We will discuss how to improve software for the humanities and the social sciences by designing programs based upon computer science principles that are at the same time grounded in the needs of humanities scholars, using only those software libraries, programming languages and paradigms that truly fit our purposes. In particular, we will first explain briefly how our Intentionally-Linked Entities (ILE) database system can be used in humanities research, social networks and temporal GIS as a replacement for relational database management systems. In ILE, relationships are represented using linked data structures rather than two-dimensional tables, following the long-term trend in other areas of computing towards replacing arrays by linked data structures. Relationships of user-defined complexity

can be used routinely, and unlike in the relational model, such relationships are favored over the use of attributes. We will discuss why using ILE instead of other databases can lead to more accurate modeling of historical knowledge, as well as an almost complete elimination of redundancy.

We will present a new software tool we are developing that will allow humanities scholars to enter and analyze data. Humanities scholars often have documents that contain numerous pieces of interrelated information that need to be analyzed to reach an understanding of the material under study. These texts, most often word processing documents, are developed over a great deal of time from primary sources, requiring a researcher many months or years to compile. It is only after this monumental task that the scholar is finally able to start piecing together the information that is key to an understanding of the research topic. These scholars typically collect data from numerous documents containing interrelated players, places, and events. We will present a tool to help gather these disparate pieces of information into a database. The user interface for this application follows the well-known computer programming paradigm of the Integrated Development Environment (IDE), whereby a project is created by collecting relevant documents (source files, resource files, and supporting data) into a single project file defined by the user. As documents are added to the project, they can be analyzed by selecting text and creating entities and assigning relationships to those entities via an ILE database. The text and matching text in other documents may then be color-coded to make it easier to keep track of the entities that have been created. The entities/ relationships will be stored in an ILE database and displayed as a tree next to the document window. We will show an early prototype of this application as well as give an overview of the fundamental components: project browser, document windows, entity/relationship trees, and the ILE database. We will also discuss applications for geographically-integrated history and its relationships to social networks and researcher-guided textual analysis.

The Ethics of Virtual Cultural Representation

Szabo, Victoria
ves4@duke.edu
Duke University

Spatial/temporal platforms for visualizing humanities data offer scholars new ways of representing the people, places, and material objects they study within a larger cultural context. These environments can be used to illuminate the lives of individual people, groups and objects over time, present archival materials in geographical or topological context, model theories of structural change over time, and annotate lived experience of the physical world in realtime. While digital maps, virtual worlds, and location-based mobile applications differ in how they represent information, each has attributes that both lend them authority and power, and potentially provoke ethical challenges. This project attempts to articulate some of those features in order to develop guidelines for ethical rich media map and virtual world construction in the humanities.

As digital humanities scholars we sometimes adopt tools created for purposes different from our own. It is our responsibility to understand the rhetorical effects of the communications strategies we adopt for our teaching and research. In the case of spatial/temporal platforms like Google Earth, the built-in assumptions about content-presentation rely on consistently abstracted information. Because humanities “data” is often heterogeneous, ambiguous, incomplete, qualitative and partial, it does not always fit well with a “vector” platform that demands specificity. The temptation to assign a point, a line, a path, a region, or a date can be powerful when assembling an archive. At the same time, the “bitmap” aspects of a rich media environment, both the panoptic stitched imagery of satellite views and street views, and the user-generated annotations, naturalize and potentially humanize they system’s affect, lending it further power as a tool for spatial-temporal representation.

To combat these totalizing tendencies, we can understand our assemblages as database driven, hypermediated historical narratives, whose organizing principles of space and time offer a thru-line for our content. Doing so requires that we show the seams, make the construction produces visible in the final product, and allow the user to deconstruct, or drill-down in our creations. The separation of presentation

and data that is a mainstay for coders is difficult to maintain in this context, but is a useful guide when thinking about display possibilities. The goal is to reveal the ways in which a map or virtual-world based presentation is a coherent, but not exclusive, narrative built up out of and illustrated by the materials at hand. Ideally the presentation layer can be added onto and changed based on new information, additional data, or other perspectives.

This approach to spatial/temporal project creation has implications for how we construct, share, and annotate digital map and virtual world based project, which will be elaborated here through three examples. Each project includes both quantitative and qualitative "data" sources, georeferenced content, and a focus on a specific historical or cultural group. While none of these projects are specifically ethnographic or documentary in focus, each raises ethical challenges for the digital media author representing a populated environment. Example Projects:

1. Multimedia Mapping: Muhuru Bay, Kenya, was undertaken by a team of faculty, postdocs, students, and community leaders. This complex project involved the ISIS Program, Duke Global Health researchers, DukeEngage, a summer enrichment program for undergraduates, and the WISER Foundation, an NGO. Some of the issues that arose here were around the use of demographic and survey data reflecting the local population, testimonials by local children on their lived experience as female students in the community, the exposure of selected aspects of the local infrastructure, including a latrine quality survey as part of a publicly accessible rich media map. This project also involved the creation of what were literally the first contemporary maps of the region showing location and sub- location boundaries of the various provinces and villages in Muhuru Bay, and well as a research study on local children's perspectives on mapping as part of study of the social geographies of AIDS and HIV.

2. "The Walltown Neighborhood History Project," is a community mapping project, involved seventh and eighth grade students in a summer camp-based experience which involved overlaying historical census data information and fire insurance maps onto a contemporary Google map of their community. This project was intended to promote technology literacy among the students, to provide a chance for local community members to learn more about their neighborhood's history and contribute their own content, and for Duke researchers to leverage an archive of historical "Digital Durham" materials in a highly accessible, public fashion. Challenges here include the public presentation of historical data in a public, georeferenced context, the presentation of historically racialized demographic language to be read in the non-scholarly context of a Google map, the production of a local and collective identity for the Walltown residents via expressive and quantitative (GIS) data, and the potential exposure of individuals on the public web, even granting their informed consent.

3. The "Virtual Haiti Project," is part of a Humanities Lab focused on Haiti and supported by Duke's Franklin Humanities Institute. Students in a class called "Representing Haiti" are asking the question of what it means to create a "Haiti Island" in Second Life, and what it is and is not appropriate to model there. We need to understand how a geo-referenced constructed environment exceeds the bounds of our intentionality, which in this case is to create a Kreyòl language learning space and virtual front end to a library of Haiti-related, downloadable resources alongside a set of coordinated maps related to visualizing the Haitian diaspora. Perhaps more fruitful than direct mirror-world strategies in-world be non-representational strategies such as psychogeographies, associative place-based meditations, and imaginative transformations or reconstructions. Yet why not also a Port-au-Prince StreetView?

A System for Referencing Personal Names through Iconography and Sharing an Authoritative Information Source for Personal Names by API

Togiya, Norio
togiya.norio@iii.u-tokyo.ac.jp
University of Tokyo

Kawashima, Takanori
t_kawa@valdes.titech.ac.jp
Tokyo Institute of Technology

1. Abstract

In this research, we constructed an ontology-based name authority file using a topic map, and then, using API and other Web services, we looked into using this with navigational systems .The name authority file was created and provided through API services. By using this service, it is possible for developing navigation and analyzing tool of Digital Cultural Heritages to have a comprehensive name list of historic personal names and artists. These API are used for Digital Cultural Heritages of digitized anthropological material, old photographs.

2. Introduction

While there have heretofore been various forms of Digital Cultural Heritage, even for digitized and stored materials it has been typical to select materials through a list or by alphabetical search. However, in situations such as for historical documents, it is preferable to be able to select materials by searching for documents that deal with related individuals, social organizations, historical events or periods. In particular, a proper noun should be controlled as authority file.

For this reason, in this study, an ontology-based name authority file —which defines the relationship between people, organizations and places—was created. And the information was stored in standardized ontology language. We adopted Topic Maps for constructing name authority file and API system providing it for digital cultural heritage.

There are some examples of ontology-based name authority file using Topic Map. The unique features of this study is visualizing ontology-based name authority file using Topic Map and providing it with API. In this paper, we will describe in detail.

3. Definition of Ontology

An important goal of this archive is to select and view each material through an understanding of the various materials and the relationship between the characteristics related to them. To achieve this, we decided to use ontology relating the various items in this project:

1. Defining concepts, characteristics, and meanings of various items
2. Aiming at systematizing various concepts and items in the world
3. Aiming at generality, which makes materials reusable, and shared knowledge
4. Describing items with rules and in a language based on certain regulations

In Greek, ontology originally means existence and has been frequently used in philosophy. However, with the development of the study of artificial intelligence and knowledge engineering since the late 20th century, it began to be used as a term referring to a semantic systematization method for various items in the world in order to facilitate its understanding by machines. Meanwhile, when the Semantic Web was proposed by Tim Berners-Lee as a more effective method to connect fragments of information on the Web through the spread of the Internet in the late 1990s, ontology was used as a term to refer to the methodology used to describe content created on the Web with more regulated rules and language.

Various types of ontologies have been developed through the course of time, such as –Upper ontology aiming at the systematization of various items in the world based on philosophical discussions, Domain ontology developed to express the edifice of knowledge in a specific domain (mainly for industrial use), and Web ontology developed to systematize information and knowledge on the Web.

“Upper ontology and Domain ontology are provided to construct concepts 1), 3), and 4) aiming at 2) defined above. Web ontology initially had the practical purpose to develop item 4) in order to define the relationship between information items on creating contents using method 1) to achieve item 3), which can be seen in the combination of the Dublin Core and RDF technologies. The ontology of CIDOC CRM –used to describe the metadata of cultural resources–seems to belong to this

type. Recent advances made in Web ontology aim at the semantic systematization of content on the Web, which was its initial objective. In this context, it may be said that it has advanced in the area of item 2) by replacing the world with the Web. However, how the system of the concepts on the Web and that of the things in the real world are unified as the former is a reflection on the latter remains to be seen.

This study systematizes item 2) for historical information with method 1) in anticipation of the union of these in the future and aims at item 3) by using item 4), which can be unified with Web ontology, etc. in the future. As for now, we are mainly focusing on implementing items 1) and 2). In particular, we designed name authority file based on the above mentioned definition on ontology.

4. Constructing Ontology Using Topic Map

4.1. Introduction about Topic Map

We adopted topic map to constructing above mentioned ontology-based name authority file. Topic maps that represent ontology use the ISO/IEC JTC1 SC34 established ISO standard, which is still being revised, but as of April 2009 some parts have not yet received JIS standardization. Topic maps are a technique for classifying, organizing and making easily visible information and knowledge, and play a role similar to the index of a book for displaying information in space. It is possible to model and process the relations between subjects in the problematic areas, relations between information resources, such as Topics, Association and Occurrence with a computer.

In order to identify and discern the subjects, topic maps have a mechanism known as PSI (Published Subject Identifier). PSI, as a subject identifier, allocates and publicizes a unique IRI (Internationalized Resource Identifier) to each subject. Placed above the indicated address of the IRI, and serving as a descriptive information resource is the Published Subject Indicator, or PSD (Published Subject Descriptor), which allows the understanding of what the subject represents. For this project, as PSI was used, subjects were not given their own name or alias; rather they could be identified and discerned by their IRI.

4.2. Creation of Topic Map

Information relating to the photographer, scholar and noble family was collected to use navigation and analyzing for old photograph. Especially for information that needs to be shared, information necessary for multiple people is gathered. Information gathered in this manner is described in the creation of topic maps.

Information gathered by multiple people is shown in table 1 under the entries of "name," "reading," "nationality," "occupation," "birth and death," "biography," "hometown," "place of residence," "attached organization / group" and "title." These items are candidates for topics when creating the topic map. After completion of the topic map, navigation will become possible based on these subjects.

The three elements of which Topic maps are composed are mainly used to express a variety of subjects (topics), express relations between topics (Associations), and to link information resources related to topics (Occurrences).

These structures are adapted as shown in Table 1. The names, various related items, and reciprocal relations of items regarding the ontology at first can be thought of as topics for topic maps. Also, for name sources, each "person" and "place" and the relation among people (Associations) is adapted. Furthermore, as links to specific information resources, "birth and death," "biography," "related URL/URI" etc., correspond to appearance (Occurrences).

In addition, it is possible to classify information stored for these (Topics), (Associations) and (Occurrences). For the first topic (Topics) various "forms" were established; Topic Types, which represents the topic "form," for (Associations), defining the relation type are Association Types and Association Role Types, and for (Occurrence), Occurrence Types expresses the type of appearance.

The specific items of these "forms" can be expressed, for Topic Types, "Country," "Person," "place name," "organization/group," and "occupation" etc.

Furthermore, Association Types are established showing various relations such as "friendship relation," "marital relation," "student/teacher relation," "workplace relation", and "filial relation," Association Role Types express the role of each topic that produces relations, such as "significant role," "brother (sister) role," "wife role," "child role," and "pupil role," etc.

Next, Occurrence Types are established, "reading of name," "reading of surname," "name in roman alphabet," "surname in roman alphabet," "personal relationship," "relationship period," "birthplace," "source," "role," etc. Based on these items information is structured.

Personal topics such as parent-child relation, fraternal relation, matrimonial relation, friendship association etc. are expressed through the lines leading into and out of the personal topics.

After creating the topic map, based on these forms queries and navigation are possible, and compared to semantic processing more meaningful searches can be conducted.

Categories of personal name authority information	Correspondence in Topic Map
Name	Topic Name
ID	SubjectID
Related URL/URI	External occurrence
Dates of birth and death	Internal occurrence
Brief biography	Internal occurrence
Place of birth (multiple responses possible)	Linked by association to other topics
Place of residence (multiple responses possible)	Linked by association to other topics
Relationship (multiple responses possible)	Association

Table 1. Major Components of the Topic Map

5. An Overview of Iconographic Analysis Using Authoritative Information.

In this study, as shown in Fig. 1, an ontology-based name authority file was constructed, and that information was distributed through API. For name authority file, focusing on photographers, scholars and noble family, roughly 3000 names were entered into a DB system. This authoritative personal name information was mainly made from standardized reference materials used in museums, libraries and archives, and created from element sets which allowed the sharing of information in this study, and API was added to the database to enable external searching. In addition, as an external system a picture annotator was created using this API.

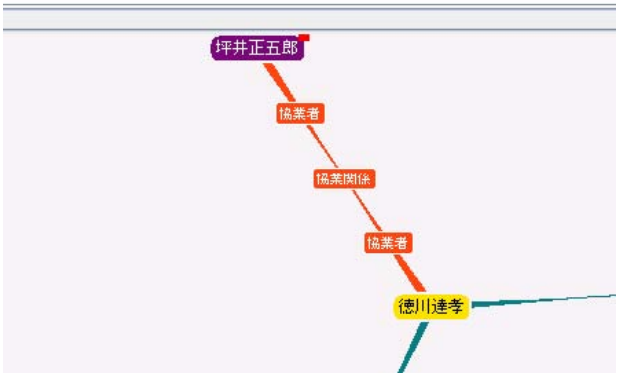


Figure. 1. Mutual Relationships among Topics

In this study, a simple API was designed to make personal information taken from a reference DB searchable. By using this API, it is possible to use even personal name information controlled by an external system.

There were four Parameters that could be used (personal name, birth year and year of death, family, personal ID) to search, and the eight types of data that could be obtained from search were (personal name, personal ID, birth year or year of death, family, biography, related individuals and their relations, and source). It was possible to identify even people and families that had a unique ID, as well as individuals who shared the same name. Furthermore, by using the controlling ID, semantic interchange with an independent external system could be made possible.

The goal of the current design is not to have all fields of the DB searchable by semantic search, but to provide those fields thought to be frequently utilized by external applications simply. If a higher order search is desired, it can be performed from the editing UI of the authoritative information DB. The XML-RPC was used for the API protocol.

6. Designing Name Autholity File for Navigation and Analysing Image

In the last section, we constructed an ontology-based name authority file using a topic map. It is possible to use these to confirm interpersonal relationships and relationships between people and organizations while looking at a graph. In this research, we constructed a name authority file concerning scholar, photographers, etc. with a topic map. Fig 1 is a graph showing these people's interpersonal relationships and their relationships. This chart mainly shows the relationships between the parties in relation to the anthropology materials. In this way, it is possible to

use a personal name information authority using topic maps for navigation systems.

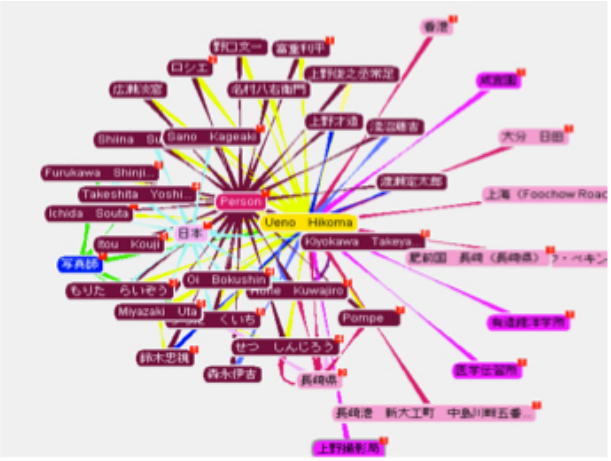


Figure. 2. Relationship among photographers

Fig 2 is a graph showing relationships of photographers. Also, as shown in Fig 3, it is possible to show detailed information regarding these personal names, etc. by clicking on the personal name icons.

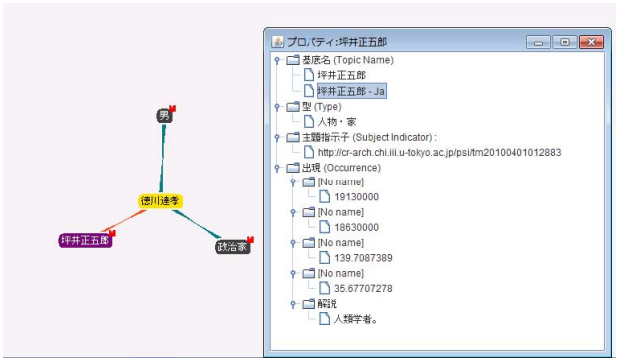


Figure. 3. Name Authority File

Furthermore, it is possible to assign annotations to the materials, as well as to assign personal name information which has been regulated by the topic maps. It is also possible to provide information to the annotations using the regulated personal name information. With this, it is possible to search through material annotation information from the personal name information that has been regulated by the topic maps.

It is possible to combine this topic mapped personal name information with the digital cultural heritage main body, but it is also possible to put it to use through API. By using API, it is possible to use this information for navigation and annotation with other digital cultural heritages. This makes it possible to regulate the information and analyze the materials using information which defines the relationships.

7. Results and Issues

As seen above, ontology-based name authority file utilizing topic map and API for Digital Cultural Heritage was explained through application. In this study, ontology-based name authority file was constructed. By using topic maps, related people, can be clarified, and materials can be viewed. And API accelerated utilization of name authority file

In addition, by allocating URI to personal topics, personal identification can be performed by URI, not just by name, and by using the merge feature it is possible to share data from other topic maps. Also, by publishing with PSI information can be shared by other RDF and topic maps. For IRI, which kinds of character strings are appropriate is a matter for future consideration.

Concerning visualization, as demonstrated in fig. 2, visualization of relationships is possible. However, this causes problems such as the difficulty in recognizing relations that appear between more than three topics to arise. This problem also occurs with other viewers, and further study is necessary.

It is also necessary to consider whether and how to share the topic map created for this project with other projects and databases. As mentioned earlier, as each topic is given a universal URI, it is necessary to consider how it is necessary to link to other topic maps.

Furthermore, for this project, previously created information was converted into a topic map, but in the future it will be necessary to design a system to handle increased topics for Digital Cultural Heritage. Along with the need to store more materials, a system to increase information is necessary.

ACKNOWLEDGMENT

I would like to give special thanks to Motomu Naito (Knowledge Synergy Inc), Hirotada Kobayashi (Cooba Corp).

References

http://www.topicmaps.org/

Kivelä, A., Lyytinen, O. (2007). 'Case study: publishing large collection of artworks using Topic Maps'. *Topic Maps Users Conference.* Oslo.

Norio TOGIYA, Akira BABA (2007). 'Constructing Integrated Digital Archive Using Ontology and User Community'. *Archives & Museum Informatics,*

ICHIM07 International Cultural Heritage Informatics Meeting Proceeding. http://www.archimuse.com/ichim07/papers/togiya/togiya.html.

Riichiro Mizoguchi (2004). 'Ontology Engineering Environments'. *Handbook on Ontologies.* SpringerPp. pp. 275-298.

http://cidoc.ics.forth.gr/

Nicolas Guarino (1997). 'Some organizing principles for a unified top-level ontology'. *Working Notes of AAAI Spring Symposium on Ontological Engineering* , Stanford, .

The Wheaton College Digital History Project: Digital Humanities and Undergraduate Research

Tomasek, Kathryn
ktomasek@wheatonma.edu
Wheaton College

1. Digital History

Can undergraduates contribute meaningfully to a long-term digital history project? What role can transcription and markup play in the undergraduate history curriculum? How can collaborations among instructor, archivist, and technologist contribute to undergraduate research? What is the role of collaborations with other small liberal arts colleges and with large research universities?

All too often, students majoring in non-science disciplines have little exposure to computational thinking and working with computer code. At the same time, digital methods of analysis exert growing influence on the practice of many disciplines in the humanities and social sciences. The Wheaton College Digital History Project seeks to bridge this gap using tools from Digital Humanities.

2. The Wheaton College Digital History Project

Since fall 2004, undergraduates in History courses at Wheaton College have been transcribing and marking up nineteenth-century documents from the Wheaton College Archives and Special Collections for digital publication. The opportunity to begin this work arose when a confluence of events combined new interest in and experience with the Text Encoding Initiative (TEI) at Wheaton College and the acquisition of the pocket diaries of Eliza Baylies Wheaton. Beginning in January 2004, Wheaton College collaborated with Mount Holyoke College to host a two-part conference that explored uses of TEI in teaching and research at liberal arts colleges. The conference included instruction in TEI from Julia Flanders and Syd Bauman of the Women Writers Project at Brown University.

Our first effort at incorporating TEI into a course occurred in fall 2004. Students in an introductory course on women's history, U.S. Women to 1869,

learned about the economic uncertainties in the lives of unmarried white women when they transcribed and marked up the journal of Maria E. Wood, the daughter of a Maine Baptist minister. Each student was assigned to transcribe a set of pages from the journal, and then groups of students marked up the entire journal using themes from the course: family, work, religion, death and mourning. At the end of the course, the students expressed a sense of having gotten to know Wood and having understood the past better than they ever had before.

In subsequent semesters and summers, students have collaborated with members of the faculty and staff to create digital editions of the diaries of Eliza B. Wheaton. The three students who worked together in summer 2005 became a community of enthusiastic historians. As they transcribed the travel journal and pocket diaries of Eliza Baylies Wheaton, they used their spare time to explore the town of Norton, especially its cemeteries, where they looked for the birth and death dates of people mentioned in the documents. During subsequent summers, student workers continued to transcribe and mark related documents.

In spring 2009, the project took a new turn, as students in the research methods course for History majors began to transcribe and mark up pages from the daybook that Laban Morey Wheaton kept between 1828 and 1859. This book records financial transactions that reflect some of the range of Wheaton's business interests during these forty years, including agricultural pursuits and rentals for land and houses as well as tax collections, fees for legal services, and the operation of a general store in Norton, Massachusetts.

3. Undergraduate Research

3.1. Teaching Module

Students examine Laban Morey Wheaton's daybook alongside his account ledger and cashbooks that date from the same period to get a fuller idea of the financial context for the daybook transactions. Each student transcribes a two-page spread using a Google spreadsheet to facilitate keeping track of the tabular data. The academic technology liaison converts these documents into XML files that students open in oXygen and mark up using the guidelines of the TEI. Students code the transactions for personal names, commodities, amounts purchased, amounts charged, and mode of payment—cash or credit.

In another class meeting, the academic technology liaison demonstrates visualization tools to show students examples of ways to display results of querying the files. Students write short History Engine episodes based on the transaction of their choice, in preparation for writing papers based on the data they have transcribed, coded, and queried.

4. Collaborations: Archiving, Pedagogy and Research

Since Wheaton College is a liberal arts college that prioritizes teaching and learning, including students in the process of scholarly research makes sense, as does promoting a collaborative pedagogy in which members of the faculty and staff come together to deploy their complementary expertise in leading students through the process of research and writing. The College Archivist and Curator of Special Collections identifies the documents to be transcribed in collaboration with the instructor, who uses her scholarly expertise to identify secondary sources that will help the students contextualize and interpret their findings. The archivist guides students to sources that help establish the local context for the documents. The Technology Liaison for Humanities instructs students in markup and querying data. He also performs backstage transformations that make the data available in appropriate forms for students to mark up, query, and manipulate.

Students contribute more than their labor since they bring to the documents a perspective distinct from those of the archivist, instructor, and technologist. Their very unfamiliarity with the historical context allows them to bring to the project questions that enable new insights into the implications of the data that they help create. Whilst we do not expect them to contribute to the digital humanities at the same level as graduate students, we do value their participation, and we hope to prepare them for advanced study in the field.

5. Next Steps: Transcribe Wheaton

Since we have developed a workflow and are comfortable with using this assignment as a teaching module, we are piloting targeted crowd-sourcing to speed transcription. Students continue to transcribe, mark up, query, and write about transactions from the daybook in courses.

Transcribe Wheaton, modeled on the Transcribe Bentham project at University College London, will provide a portal and tools to allow students and

friends of Wheaton College to participate in digitization of financial documents from the Wheaton Family Papers either as part of their coursework or as a volunteer contribution to the ongoing work of the Wheaton College Digital History Project. We hope that faculty members in courses in Anthropology, Computer Science, Economics, and other fields will use the teaching module to employ transcription and markup in their courses. We also hope that graduates of the college will assist in the project, and we expect to be able to open transcription beyond the college by 2012.

6. The Long Term

When we began transcribing the pocket diaries of Eliza Baylies Wheaton, we chose TEI because we appreciated the flexibility of XML and we liked the idea of performing transcription once and being able to manipulate the data thereafter. That initial decision has affected the choices we made as we began transcribing the financial records, which are much more abundant in the Wheaton Family Papers collection than are diaries and letters. Such records are in fact abundant in many archives, yet they are underutilized by historians, in part because of their inaccessibility. We hope that our project will contribute to the development of standards for TEI markup of tabular records, thus encouraging similar projects that will increase the accessibility of historical financial records and other tabular data.

Since our project combines archival, pedagogical, and scholarly purposes, we find ourselves negotiating constantly among them. The collaborative nature of our project and the multiple varieties of expertise and interest brought to it by each member of the team necessitate ongoing consultations among the archivist, instructor, and technologists. Like many projects in digital humanities, ours requires a certain amount of comfort with technology from the archivist and instructor, combined with an equivalent amount of comfort with the humanities from the technologists.

Because ours is a small liberal arts college, we rely on collaborations with other institutions to support our project. Our technologists, for example, have spent 2010 collaborating with colleagues from other liberal arts colleges as well as from Brown University and the University of Virginia as they plan a presentation tool for TEI documents produced at small colleges. And members of our project team have taken additional courses with Julia Flanders and Syd Bauman as we continue to hone our TEI skills.

Our poster/demo features our project website: http://wheatoncollege.edu/dhp/. Whilst the website currently serves as a brochure for the project, we hope to include additional features by the time of the conference, including links to student writing about individual transactions from the daybook, as well as a portal for transcribers.

We share our project as an example of a long-term project in digital history that includes undergraduates as significant partners in the digitization and interpretation of a hidden collection that offers insight into the relationship between capitalist accumulation and women's education in the nineteenth-century United States.

"The Start of a New Chapter": Serialization and the 19th-Century Novel

Truxaw, Ellen
ellen.truxaw@gmail.com
Stanford University

Although there have been numerous efforts to theorize the novel, little work exists on the chapter. Previous work has focused on the chapter as a device to structure the internal events of a narrative, as well as a means to affect narrative pace. In his essay “The Chapter in Fiction” Philip Stevick discusses the chapter as a structural unit containing events; however, common critical assumptions have been divided over whether to treat the chapter as a bibliographical or formal unit. This study identifies the chapter as a formal unit that is influenced by historical and bibliographic pressures. The chapter's ubiquity and seeming formlessness have made it a difficult object of study with traditional tools of literary analysis, but techniques in humanities computing provide new access to the stylistics of the chapter.

In his book, *The Sense of an Ending*, Frank Kermode describes novels as “fictive models of the temporal world” in which readers must find a balance between realistic representations of time and necessary narrative deviations from this chronicity (Kermode 54-55). My stylistic analysis of the chapter focuses on grammatical and semantic temporal markers at the beginnings and ends of chapters. By investigating the stylistic trends in 19th century authors’ treatment of narrative time, I argue that historical, literary-historical, and bibliographic forces shape the form and function of the chapter across 19th century.

By the Nineteenth Century, the novel had been established as a largely chapter-based form; however, pressures of serialized publication demanded that authors reconsider the way they structured both their narratives and chapters. In a letter to Elizabeth Gaskell, Charles Dickens offers advice on how to write successful material for serial publication. He writes about what would happen if text that is not intended for serial publication gets divided and serialized anyway. He writes, "The scheme of the chapters, the manner of introducing the people, the progress of the interest, the places in which the principal places fall, are hopelessly against it. It would seem as though the story were never coming, and hardly ever moving. There must be a special design to overcome that specially trying mode of publication" (Grubb 143). Dickens insists that serial publication required novelists to rethink the way they structure their narratives. This study analyzes both serialized and non-serialized 19th century novels to ascertain the extent to which serialization affected the chapter as a formal unit.

Using distant reading techniques, I have studied stylistic trends in thirty-six Nineteenth Century novels: Non-serialized -- *Pride and Prejudice*, *Persuasion*, *Frankenstein*, *Ivanhoe*, *The Entail* (1823), *The Last Man*, *Pelham* (1828), *The Heir of Redclyffe* (1853), *Barchester Towers* (1856*), Jane Eyre*, *Villette*, *Henry Esmond, Tancred* (1847), *Mary Barton* (1849) *Mill on the Floss* (1860), *The Coral Island* (1858), *Black Beauty* (1877), *Dracula* (1897) Serialized-- *Jack Sheppard*, *Barnaby Rudge*, *David Copperfield, Vanity Fair, Hard Times, North and South, Our Mutual Friend*, *Far From the Madding Crowd*, *The Woman in White*, *The Moonstone*, *Middlemarch*, *Beauchamp's Career*, *Can You Forgive Her?, The Way We Live Now*, *Wives and Daughters*, *David Copperfield*, *Treasure Island*, *The Trumpet Major*. I selected novels based on their canonicity and tried to study works from a wide historical range. By close-reading chapter beginnings and endings, I have created classification categories for the main "types" of beginning and endings. These "types" treat diegetic temporality in distinct ways through different verb tenses and deictic markers. I classified chapter beginnings into the following types:

- Type 0 — The narrator describes a character or place without direct references to diegetic time.
- Type 1 — The narrator starts the chapter on "the next day"
- Type 2 — The narrator begins the chapter where he leaves off, or *in medias res* often using the past progressive
- Type 3 — The narrator summarizes the events of an interim period before entering a scene in the chapter.
- Type 4 — The narrator uses phrases like "one day" to place the reader in an undefined place within diegetic time.

I differentiate these types based on defining grammatical and semantic markers that indicate the author’s different representations of temporality. I read the beginnings of every chapter from each of the novels and classify them into the category into which they fit best. I then compare the relative frequencies of types of chapter beginnings and endings using $\chi 2$ tests to determine whether significant differences between

serialized and non-serialized novels appeared. These differences proved significant. Notably, the Type 3 and 4 beginnings occurred significantly more frequently in non-serialized novels than serialized novels. The difference in the relative frequencies of beginning types in serialized and non-serialized novels suggests not only that the chapter exists as a formal unit of narrative, but that historical and bibliographic changes affect this form.

Collaborating with Matthew Jockers and the Stanford Literary Lab, I am continuing to look at trends in a wider data set of novels. Matt Jockers has built software with the primary purpose of classifying the beginning and ending paragraphs of chapters in novels from the *Chadwyck Healey 19th century British Fiction* database. By obtaining data on over 250 novels, I aim to investigate the prevalence of particular beginning and end types across the 19th century as a whole to elucidate how and why authors divide their narratives as they do as well as to understand the role narrative time plays in these divisions.

References

Grub, Gerald Giles (1942). 'Dickens' Pattern of Weekly Serialization'. *ELH.* 9:2: 141-156.

Kermode, Frank (1966). *The Sense of an Ending.* London: Oxford University Press.

Stevick, Philip (1970). *The Chapter in Fiction; Theories of Narrative Division.* Syracuse, N.Y.: Syracuse Univ. Press.

Adapting EATS for Crowdsourcing: Register Medicorum Medii Aevi

Viglianti, Raffaele
raffaele.viglianti@kcl.ac.uk
King's College London, United Kingdom

The Register Medicorum Medii Aevi (RMMA) is an active pilot project at King's College London that seeks to lay the foundations for an interdisciplinary online register of doctors in the middle ages. It is common for similar kinds of historical online resources to prepare a database out of a large collection of scholarly findings. RMMA, instead, aims to create a growing database which is populated over time by various scholars and students to make their findings accessible on the web. In order to pursue this task, the prototype for this pilot must allow data insertion from a wider group than the project partners alone; therefore several crowdsourcing principles are applied and extensive user testing is being undertaken. This poster will show solutions to the main challenges encountered in the design of the database and in adapting EATS, a Django application for authority records, for wider online use.

2. Overview

Researchers from different areas of medieval studies are likely to encounter information in primary sources about *medici*, the physicians of the time. Information about *medici* and their practices was recorded in a variety of documents; they would often travel across cultures, which can serve as indication of their high value. This has produced many sources from different cultures and in different languages. Because different areas of study often gravitate around geographical, cultural and linguistic divisions, it is challenging to investigate the movements of people and knowledge across these barriers. RMMA explores how digital publication can help in exchanging reliable information between researchers in such areas and currently involves experts on Anglo-Saxon, Arabic, Armenian, Greek, Jewish, Latin materials that provide information about doctors, and medical knowledge in the medieval period.

To address this issue, the project is developing a web application that will allow to (1) insert information about *medici* from primary and secondary sources; (2)

handle different opinions about the same individuals; and (3) connect to prosopographical projects that deal with a similar historical time-frame and with historical geographical resources. [1]

3. Technologies Involved

The application is built using Django,[2] a Python open source web application framework aimed at simplifying development through a reusable set of common libraries and components. Developers can code their own application to be re-used within other Django frameworks, which is a factor that contributes towards the web framework's success. EATS (Entity Authority Tool Set) is a Django application developed by Jamie Norrish at the New Zealand Electronic Text Centre (NZETC) and "is now used to manage the 30,000+ entities represented in the NZETC online collection of significant New Zealand and Pacific Island texts" (Norrish 2007). The application targets issues usually faced by libraries, such as ambiguous identifiers based on names and titles. By improving on traditional authority systems where one identifier groups different entries describing the same entity, EATS allows the linking together of different collections and authorities, introduces user management and provides greater scope for disambiguation. RMMA found that EATS's approach could greatly contribute to the objectives of the project's web application, specifically in handling information about the same entities coming from multiple sources.

Most EATS top-level components have been adapted for RMMA's purposes: Entities are database tables that must be associated to one or more Authorities. Authorities may then make Property Assertions that associate Entities to specific default or user-defined properties. In RMMA, Entities are medici, patients, employers, locations, institutions, etc., while Authorities are the contributors of the data, such as authors of narrative secondary sources or contributors inserting information anew from a primary source (note that in this case the authority remains the contributor and not the primary source; in theory, different contributors can enter different information from the same primary source, intentionally leaving room for debate). Property Assertions in RMMA reflect statements such as: Authority *contributor* states that Entity *medicus* has name *name*, or similar. Finally, users of the system are able to insert data for one or more authorities, which is particularly useful if one user is entering information from a number of secondary sources. When users enter data discovered by themselves, however, they are acting like authorities. Even in this case, for the purpose of the system, the dichotomy between user and authority must be maintained: the scholar will insert data as a certain user controlling the authority that represents him or herself.

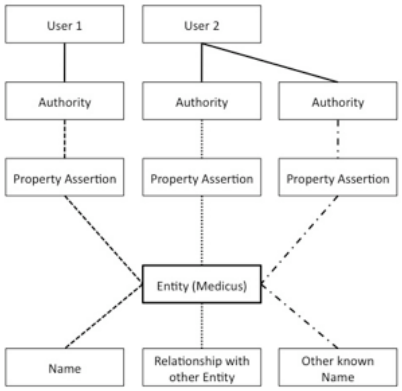


Fig 1. This graph shows the role of users in controlling authorities and different authorities declaring statements (property assertions) about the same entity.

4. Adapting EATS Interface for a Wider Audience

The default EATS interface for data entry is designed for completeness and may require some training for the inexperienced users. After all, this interface is likely to be used by personnel for populating the database and most of the design effort is better spent on the querying interface that end users will see. In the case of RMMA the project is vastly adapting the interface, as data entry will be an essential part of the end user experience. The project has been looking at other initiatives in the Digital Humanities and Digital Library fields that are adopting crowdsourcing for collecting data.[3] Crowdsourcing consists of outsourcing specific tasks to be performed by volunteers via a web application. The term is a neologism coined in 2006 in an article on Wired, in which the author claims that the current pervasiveness of technology has reduced the gap between professionals and amateurs. Rose Holley (2010) argues that crowdsourcing could be of great value for Digital Libraries, especially for accomplishing goals that would require a large number of staff, time and resources. In order to attract contributions from volunteers it is necessary to follow certain principles when designing the web application that they will use. Holley outlines some "tips for crowdsourcing" out of her personal experience and by analysing successful projects.

RMMA is not strictly a crowdsourcing project, as submissions will be reviewed by a board before being published. The reason for this is twofold: on one

hand it is desirable to control what gets into the database, though successful crowdsourcing initiatives usually found volunteers good at policing each other and keeping away malicious users; on the other hand, RMMA attempts to promote the use of the database as a publishing environment for scholars. The presence of a board, therefore, offers a model closer to peer-reviewed publishing, which is not a novel approach, as some academic projects have already been experimenting along similar lines.[4]

Despite these plans, which move RMMA away from the definition of crowdsourcing, many of the principles outlined by Holley are of great importance. Specifically, she suggests having a clear and ambitious goal, acknowledging contributors, reporting on the progress of the resource and making the application easy, reliable, quick and intuitive. Current efforts focus on improving the application by simplifying the interface and making clear paths for data entry depending on the nature of information that the users intend to store. To insure that the interface design is effective, several stages of user testing are planned.[5] RMMA is organising a number of workshops in Europe and in the United States as part of the pilot project. This is intended to reach the desired audience and at the same time have the user test the resource and collect feedback during the following few weeks.

5. The Future

RMMA hopes to use its web application to involve middle ages medical researchers from different fields separated by specific geographical and linguistic studies. The database aims to be both a growing collection compiled by interested participants and a reliable publication of information about relevant individuals from the middle age worlds. It is therefore necessary to include crowdsourcing principles in the design of the web application to achieve the proposed goals. While not a stated aim of the project, the possibility of tracing of the movement across cultures of individuals conveying and acquiring medical knowledge would be a desirable and likely outcome of far-reaching historical interest.

References

Finkel, R. et al. (2007). *The Suda On Line.* `http://www.stoa.org/sol/about.shtml` (accessed 31 November 2010).

Holley, R. (2010). 'Crowdsourcing: How and Why Should Libraries Do It?'. *D-Lib Magazine.* 3/4.

Norman, Donald A. (1988). *The design of everyday things.* Cambridge (MA, USA): MIT Press.

Norrish, J. (2007). 'EATS: an Entity Authority Tool Set'. *Australia New Zealand Digital Encyclopedias Group Meeting.* Sydney, Australia, 7-8 December.

Norrish, J. and Stevenson, A. (2008). 'Topic Maps and Entity Authority Records: an Effective Cyber Infrastructure for Digital Humanities'. *Digital Humanities 2008.* Oulu Finland, 25-29 June.

Terras, M. (2010). 'Crowdsourcing Cultural Heritage: UCL's Transcribe Bentham Project'. *Seeing Is Believing: New Technologies For Cultural Heritage.* International Society for Knowledge Organization. University College London, London, 9 June 2010.

Notes

1. RMMA is looking in particular at prosopographical projects developed at the Centre for Computing in the Humanities at King's College London, including the Prosopography of the Byzantine Empire and Prosopography of the Byzantine World <http://www.pbw.kcl.ac.uk/>. For historical locations, we are working towards integrating a query system to Pleiades when entering new locations <http://pleiades.stoa.org/>.
2. http://www.djangoproject.com/
3. See in particular: Transcribe Bentham, a crowsourcing project hosted at the Centre for Digital Humanities, University College London <http://www.ucl.ac.uk/transcribe-bentham>; the image collection of the Victoria & Albert Museum <http://collections.vam.ac.uk/crowdsourcing>; and other projects discussed in Holley 2010.
4. An exemplary project is Suda on Line: Byzantine Lexicography (SOL) <http://www.stoa.org/sol>, which involves more than a hundred scholars for the translation and editing of the *Suda*, a 10th century CE Byzantine encyclopedia. The large size of *Suda* (30,000+ entries) has encouraged the involvement of a large number of collaborators to work online. Submissions to the database are made immediately available to the public, but SOL implements a colour code that differentiate just submitted entries and entries that have been reviewed and approved by a the editors overtime. This mixed system simplifies submission whilst allowing a control of the quality of the publication.
5. As recommended by many studies on User Centred Design, after Norman 1988.

Computational Discovery and Visualization of the Underlying Semantic Structure of Complicated Historical and Literary Corpora

Walsh, John A.
jawalsh@indiana.edu
Indiana University

Hooper, Wally
whooper@indiana.edu
Indiana University

This admirable invention [computation by logarithms] added to the ingenious algorithm of the Indians, by reducing to a few days the labour of several months, doubles—if we may so speak—the life of astronomers, and spares them the errors and disgust inseparable from long calculations . . .

–Pierre Simon de LaPlace (1749-1827) Système du Monde, liv. iv, chap. iv

In addition to offering the possibility of new forms of scholarship beyond the traditional article or monograph, digital humanities and the computational tools developed as part of many digital humanities projects allow scholars to practice an accelerated hermeneutics and analysis on ever larger collections of texts and other cultural artifacts. Most projects that create scholarly digital editions of important corpora are consciously addressing a range of research issues in their respective fields and disciplines. Many of these projects are now also developing components and applying computational tools to exploit their new resources with the goal of solving those problems in an efficient and effective way.

Napier's logarithms (1614) eased the efforts of early modern astronomers. Today, computational tools promise to accomplish lengthy mechanical tasks of review and notation that generations of scholars have done by hand. Effective methods now exist that will aid trained professionals to see into the heart and structure of digital corpora.

Our project is associated with two major digital editions based at Indiana University: the Chymistry of Isaac Newton Project (http://www.chymistry.org), led by William R. Newman, general editor and co-investigator, which is publishing a digital edition of one hundred nineteen alchemical manuscripts written by Isaac Newton, thirty-two of which are now online; and the Swinburne Project (http://www.swinburneproject.org/), edited by Walsh, which is publishing the works of Victorian poet and critic Algernon Charles Swinburne.

The National Science Foundation has funded a three-year project (2009–12, #0620868) to develop computational tools for the analysis of the alchemical language in Newton alchemical corpus. We are investigating the usefulness of methods from computational linguistics, information retrieval, and network sciences for the analysis of the contents of the Newton corpus and the visualization of its structures. We have been running experiments on IU's Teragrid supercomputer system and at the Computational Linguistics Lab for the last two years.

We have developed a working suite of analytical tools that support effective real-time investigation of the semantics and structure of the corpus. Data from our analytical tools is fed into another suite of tools, the Network Workbench (http://nwb.slis.indiana.edu/), to produce meaningful visualizations that help the user to comprehend the breadth of the materials and their interconnections.

Newman has already used the tools to demonstrate the order of composition of significant parts of Portsmouth MSS. 3973 and 3975, two of Newton's most important notebooks recording his alchemical experiments performed at Cambridge.

While the tools are being designed primarily for the analysis of alchemical language in Newton's alchemical corpus, we of course hope the tools will be useful in analyzing documents from other disciplines, and so we are also applying these tools to a significant corpus of literary documents by Swinburne.

Among Swinburne's works are six volumes of collected poems, five volumes of collected tragedies, and a number of volumes of prose literary criticism. The style, vocabulary and structures of Swinburne's 19th-century literary works are obviously quite different from Newton's 17th-century scientific texts. Furthermore, Swinburne's prose criticism is very distinct—in style, vocabulary, and structure—from his poetry. An earlier effort from The Swinburne Project, reported on at last year's conference, focused on thematic networks in Swinburne's 1880 volume Songs of the Springtides. Many instances of networked themes were found by simple keyword searching. More elusive instances were identified by traditional methods of close reading,

study, and analysis. We are exploring whether our computational tools are successful at locating some of these more elusive instances of previously identified themes and whether the tools can identify clusters that suggest additional themes and streams of meaning.

Our poster and real-time demonstration of the tools will present our results thus far with the Newton and Swinburne projects.

2. Methods and Results

We have found that Latent Semantic Analysis (LSA) is very effective for discovering the semantic structure and organization of Newton's alchemical corpus. LSA methods use singular-value decomposition (SVD) to uncover hidden layers of structure in a body of information and make it possible to capture and expose similarities and dissimilarities across an entire corpus with a single set of operations.

LSA technologies were discussed in the literature in the early 1990s primarily in the context of information retrieval. That was a problem because LSA methods for formulating and executing freeform user queries over growing datasets are less effective than other methods based on Bayesian modeling like Latent Dirichlet Allocation (LDA), Topic Analysis, and Latent Profile Analysis, which superseded LSA after 2003. Many who follow the current literature regard LSA as an old and limited technology. But the goals of close study and analysis are often distinct from the goals of information retrieval, and there is little useful literature on the use of SVD and LSA for actual text analysis beyond a report produced by a team at Sandia National Laboratories in September 2010.

For our purposes, methods like LDA and LPA, which are based on identifying k-dimensional patterns from one set of document to the next and strategically neglecting what is not pertinent to those dimensions, are less useful for text analysis than methods that explicitly capture and model all relations between all words and document chunks in corpora. SVD and LSA do that very well.

Our research plan includes collocational analysis of the Newton manuscripts (almost a millions word of English, Latin, and French), and we are collaborating with IU computational linguist Sandra Kübler to do part-of-speech markup, lemmatization, and sentence parsing. As part of that effort, we have been varying the size of our LSA document chunks downward from 1000 words to 250 and 100, approaching the sentence level where the distinction between "bag-of- words" approaches and syntax-based methods tends to vanish for practical purposes and the two can be welded in combined operations.

We have created web-based components that generate comprehensive lists of correlated results based upon user-selected documents, document fragments, or word. In the case of document-fragment queries, the user can select any pair of correlated fragments and view them side by side. (See Figure 1)

Figure 1: Screen capture of user component showing a side-by-side view of two brief fragments from Portsmouth Add. MS. 3973, f. 19r, and Add. MS. 3975, f.77r: The yellow color and red dots indicate shared vocabulary and passages.

Strongly correlated fragments (those correlating at 0.7 and above on a scale ranging between -1.0 and 1.0) almost always share long phrases and sentences. The method is robust enough to detect paraphrases too. One SVD calculation detects and maps all the passages in that large, arcane corpus which share vocabulary, phrases, and ideas, saving investigators considerable time in working through those texts and revealing correlations that might otherwise escape notice.

A useful way to present SVD results for interpretation is to graph them as networks—the document chunks or the terms are graphed as nodes and the cosine similarities or correlations between them as edges. Our tools make network graph files for use in the Network Workbench (NWB) tool developed by our colleague Katy Börner's team at Indiana University. (See Figure 2)

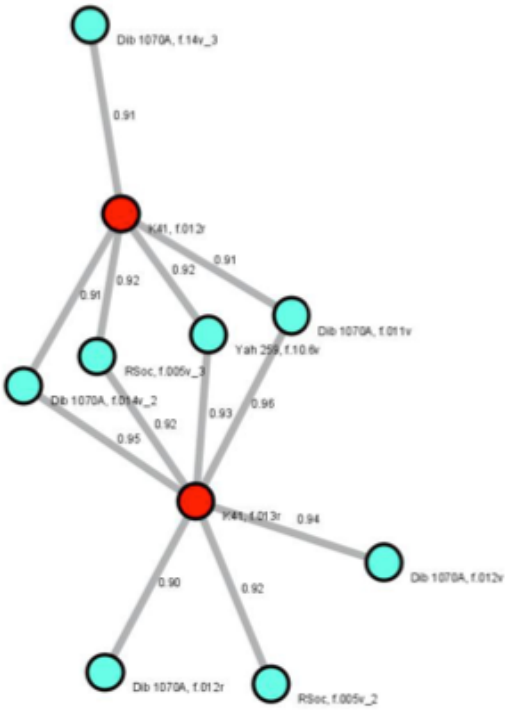


Figure 2: Sample network graph of relations between a small number of correlated fragments in Keynes MS. 41, Dibner MS. 1070A, Yahuda MS. 259, Babson MS. 417, and Royal Society M/M/6/5.

NWB is designed to be a "large-scale network analysis, modeling and visualization toolkit for biomedical, social science and physics research." Another of our contributions is testing this tool in humanities research contexts for which it was not initially designed.

Figure 2 shows that two passages on Keynes 41, ff. 12r and 13r, are better correlated with four fragments in other manuscripts than they are with each other (their correlation is 0.46). Newton worked back and forth between sets of reading notes as he attempted to make sense of the alchemical literature. The algorithms and graphs make all the connections visible at a glance.

We expect that the combination of LSA-based tools and network graphs derived from them will provide invaluable means of problem solving and discovery for the experienced researcher who is working with large, complicated corpora.

References

Börner, K. (2007). 'Making Sense of Mankind's Scholarly Knowledge and Expertise: Collecting, Interlinking, and Organizing What We Know and Different Approaches to Mapping (Network) Science'. *Environment and Planning B: Planning and Design.* Pion 5: 808-825.

Deerwester, S. C., Dumais, S. T., Landauer, T.K., Furnas, G. W., Harshman, R. A (1990). '"Indexing by latent semantic analysis."'. *Journal of the American Society for Information Science.* 6: 391–407.

Dunlavy, D., Shead, T.M., Crossno, P.J., Stanton, E.T. (September 2010). '"ParaText - Scalable Solutions for Processing and Searching Very Large Document Collections: Final LDRD Report."'. 'Technical Report SAND2010-6269'. Albuquerque, NM and Livermore, CA: Sandia National Laboratories.

Skillicorn, D. (2007). *Understanding Complex Datasets. Data Mining with Matrix Decompositions.* Boca Raton, FL: Chapman and Hall.

Walsh, J. (2010). '"Quivering web of living thought:" Conceptual networks in Swinburne's Songs of the Springtides.'. *A. C. Swinburne and the singing word: New perspectives on the mature work.* Y. Levin (ed.). Farnham, England: Ashgate, pp. 29-53.

UCLDH: Big Tent Digital Humanities in Practice

Warwick, Claire
c.warwick@ucl.ac.uk
King's College London, United Kingdom

Mahony, Simon
s.mahony@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Nyhan, Julianne
j.nyhan@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Ross, Claire
claire.ross@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Terras, Melissa
m.terras@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Tiedau, Ulrich
u.tiedau@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Welsh, Anne
a.welsh@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

There has been a great deal of concern recently about questions of how we should define Digital Humanities. John Unsworth (2010) in his plenary lecture at DHSI asked how we might define the boundaries of our discipline. UCL's own Melissa Terras, (2010) in her widely reported plenary at DH2010, warned us that we must not only understand our discipline ourselves, but be able to communicate it succinctly to others. Others, including one of the authors of this proposal, tend to the view of 'more hack less yack'. (Meloni, 2010) Yet questions remain, and the theme of DH2011 clearly prompts us toward such considerations. As a result we present a proposal below for a poster about the establishment of the new UCL Centre for Digital Humanities, (UCLDH) (http://www.ucl.ac.uk/dh/) one of whose founding principles is that of inclusivity, interdisciplinary and the broadest sense of definition, in which we demonstrate ways in which the big tent attitude to digital humanities is put into practice. Our tent includes not only other disciplines within academia, but also libraries, museums, archives, cultural heritage practice and commercial information providers. In the following proposal we discuss how this has come about and justify our belief in broadly defined Digital Humanities (DH).

UCLDH does not think of itself as a DH centre in the conventional form, where anyone working on DH must come and work in one central facility. Instead, it is built on a hypertextual metaphor: it is the hub of a network, bringing together work being done in different departments and research centres within UCL and beyond. This is one of the reasons for our inclusive philosophy. We do not believe it is for us to tell people whether they are doing DH, as we conceive of it. If they would like to become part of our network, then we welcome their involvement, since we believe that exciting new research can be created from synergies in such a network, and by unexpected collaborations between disciplines. To this end we run various different networking events such as Digital Excursions. These visits to different parts of UCL and other cultural heritage institutions such as the British Library allow participants to find out about research and digital facilities they might never previously have been aware of, and to meet and talk to others whom they might never have come across. Connections created by these meetings may take DH forward in ways we cannot predict, let alone define.

UCL has unique assets as a basis for Digital Humanities research in the form of Museums and Collections and Library Special Collections, and digital art work being produced at the Slade School of Art. We are also fortunate that our location in central London means that we are close to many of the world's greatest Libraries, Museums and galleries. As a result one of the main directions in which UCLDH has sought to extend the definition of what DH might be is in working with cultural heritage and memory institutions. For example we are working with the British Museum, the Museum of London, the Victoria and Albert Museum on a project that will help us better to understand the needs and behaviours of users of digital museum objects. We have doctoral students undertaking research at various institutions, including: the British Library, to look at the use of their mass digitisation projects; The London Metropolitan Archive, where image processing will be used to try and decipher the Grand Parchment which is too damaged and deteriorated to read; The Science Museum, where the use of 3D scans of museum objects will be evaluated by the general public; and the British Museum, where work will be done on curatorial documentation of 3D scans to investigate standards and protocols for 3D capture of artefacts.

UCL has world leading research in both humanities, computer science and engineering, and we believe that as a result it is vital to engage all parts of the university's research base equally in the DH endeavour. We aim to create new knowledge both in computer science and engineering and in the humanities, as part of the same research projects. We think of computer scientists as equal research partners in our work. Computing is not conceived of as existing to provide a service to facilitate humanities research. Thus DH research takes place in the Department of Computer Science as often as in the faculty of Arts and Humanities. One project led by one of UCLDH's associate directors, aims to develop algorithms to reconstruct the Minoan wall paintings of ancient Thera. This will lead to advances in computational methods, but it also aims to redefine the existing conservation and assembly process, helping archaeologists to create reconstructions of the frescoes, and to study them in ways that would previously have been impossible.

We also believe in engaging with the users of digital resources, whether they are in academia, cultural heritage, or the broader interested public. This is the biggest possible tent that we might pitch for DH. We are highly engaged with social media in our own work, as evidenced by the UCLDH blog, and out Twitter presence (#UCLDH). However, beyond this, several of our research projects involve social networking or crowd sourcing, and aim to engage the public well beyond academia with their heritage. Transcribe Bentham (http://www.ucl.ac.uk/transcribe-bentham/) allows users to access digital copies of Jeremy Betham's original letters, to learn about the intricacies of transcribing primary sources, and then to contribute transcribed text back to the digital collection. The QRator project will use QR codes to allow museum visitors to contribute their interpretation of objects to digital interactive labels using a smart phone app. This means that crowd sourced understanding of museum objects can complement the once monolithic curatorial interpretation of what visitors ought to be seeing.

Stretching the tent even more widely, UCLDH has also caught the imagination of the wider DH and cultural heritage community internationally with its successful discussion group. Decoding Digital Humanities (DDH) This is an informal discussion group about DH established and organised by research students and staff from UCLDH. It meets monthly and is attended by students, researchers and cultural heritage practitioners from London and the south of England as well as those from UCL. It also has five new international chapters: two in Australia, and in the USA, Belgium and Portugal.

Our definition of the big, interdisciplinary tent also includes teaching and learning. Our new Masters (http://www.ucl.ac.uk/dh/courses/mamsc) will be a highly innovative interdisciplinary programme: the first in the world to have a dual designation of MA and MSc, reflecting once again our sense of the dual balance of our field. Its diverse choice of options from a wide range of disciplines responds to the complex nature of DH, including modules from engineering, computer science, geography, archaeology, anthropology, architectural studies, psychology and information studies as well as pure humanities. It also reflects the needs of the students, the skills required for a new generation of scholars as well as those wishing to pursue a career outside academia. We will also release a substantial amount of the core materials as open access digital learning objects as part of the JISC Open Educational Resources programme: further evidence of a commitment to openness and broad public engagement in teaching as well as research.

The guiding principles of our approach to DH are predicated on welcoming the sense of a field that is growing and in flux. We do not want to put up fences, and create definitions of arcane knowledge which initiates must possess to be part of our exclusive club. We wish to open wide the doors of this amazingly diverse discipline to any and all of those who would like to take part. We believe that DH should create new knowledge in both parts of the equation, of digital technologies and humanities scholarship. We believe that DH should embrace memory institutions and cultural heritage. We believe that DH should involve those who use digital resources, allowing them to contribute their ideas and content to resources, as well as being consulted about their design. But ultimately, to be true to our principles, we believe that is it not our task to define DH at UCLDH. In the spirit of social media, we propose that the definition of the field should be allowed to develop organically, taking into account the views and input of those who participate in it, within and beyond the academy. Our view of DH is crowd sourced, inclusive and ever growing: big tent Digital Humanities in practice.

References

Meloni, J (2010). 'Reporting from 'Academic Summer Camp': the Digital Humanities Summer Institute. ProfHacker, June 10,

2010'. http://chronicle.com/blogs/profhacker/reporting-from-academic-summer-camp-the-digital-humanities-summer-institute/24672.

Terras, M. (2010). 'DH2010 Plenary: Present, Not Voting: Digital Humanities in the Panopticon Digital Humanities, 2010, Kings College London'. http://melissaterras.blogspot.com/2010/07/dh2010-plenary-present-not-voting.html.

Unsworth, J. (2010). 'The State of Digital Humanities, Digital Humanities Summer Institute, University of Victoria, Canada, June 2010'. http://www3.isrl.illinois.edu/~unsworth/state.of.dh.DHSI.pdf.

BrailleSC.org: Applying Universal Design Principles to a Digital Humanities Project

Williams, George H.
gwilliams@uscupstate.edu
English, University of South Carolina Upstate

Bohon, Cory
bohon@email.uscupstate.edu
Computer and Information Systems, University of South Carolina Upstate University of South Carolina Upstate

Over the last several decades, scholars have developed standards for how best to create, organize, present, and preserve digital information so that future generations of teachers, students, scholars, and librarians may still use it. What has remained neglected for the most part, however, are the needs of disabled end-users, especially those whose vision is impaired. While professionals working in educational technology and commercial web design have made significant progress in meeting the needs of such users, the humanities scholars creating digital projects all too often fail to take these needs into account. This situation would be much improved if more of projects embraced the concept of universal design, the idea that we should always keep the largest possible audience in mind in our design decisions, ensuring that our final product serves the needs of those with disabilities as well as those without.

We are proposing a poster session to demonstrate our work in progress on BrailleSC.org, an online scholarly resource dedicated to braille and braille literacy in South Carolina. The earliest stages of development were funded by a grant from the U.S. Department of Special Education, and the site is currently supported by a Level 1 Start-Up Grant from the National Endowment for the Humanities Office of Digital Humanities. The content is managed with both WordPress and Omeka, and for each of these CMSes we are developing accessibility plugins designed to meet the needs of visually impaired users. Our hope is that these plugins become widely adopted by other digital humanities projects that use these same CMSes. During our session we would demonstrate our site, argue for the importance of accessibility, and allow audience members to experience our site in the ways a visually impaired end-user would. We will also discuss

the work we are undertaking—at the moment, very preliminary—to develop a tool that would automatically transform an alphabetic text file (encoded in HTML or TEI standards) into a well-formatted contracted braille file. We're currently looking into whether or not the WordPress plugin Anthologize might be a useful starting point for developing such a capability.

BrailleSC seeks to combine digital humanities expertise with the important insights of disability studies in the humanities, an interdisciplinary field that considers disability "a way of interpreting human differences," in the words of one prominent scholar. Digital knowledge tools that assume all end-users approach information with the same abilities risk excluding a large population of people. If the digital humanities is to accomplish the admirable goal of creating a "big tent," welcoming a diverse array of participants, then we must broaden our understanding of the ways in which these participants access digital resources. For example, visually-impaired users take advantage of digital technologies for "accessibility" that (with their oral/aural and tactile interfaces) are fascinatingly different than the standard monitor-keyboard-mouse combination, forcing us to rethink our embodied relationship to data. Learning to create scholarly digital archives that take into account these human differences is a necessary task no one has yet undertaken.

In partnership with the Center for Digital Humanities in Columbia, South Carolina, and with guidance from George Mason University's Center for History and New Media, BrailleSC aims to model the ways in which digital humanities projects can be designed and implemented with the needs of all users, regardless of disability. Users with visual impairment often access digital information through a variety of alternatives, not primarily using traditional visual cues presented from the standard graphical user interface. For example, many such users navigate information by listening to a synthesized voice reading textual material aloud to them. The software that generates such a voice is known as a "screen reader." To make navigation easier for these users, our "Access Keys" plug-in for Omeka allows users to get from page to page and section to section by pressing an easy-to-remember combination of keys. Other users require text enlargement , and our Omeka "Text Zoom" plug-in changes the size of the text to suit their needs. Future work will refine these existing plug-ins and develop additional ones for users to customize such elements as color and contrast.

As development of the site content and interface continues, we are conducting various user-testing sessions involving a diverse group of people with varying degrees of visual ability or impairment. All interface tools developed for the *BrailleSC* project have been or will be released as open source code. All content is being made available under a Creative Commons Attribution-Noncommercial-Share Alike license. Easy-to-follow instructions for how to implement the accessibility features are currently being created. The Center for Digital Humanities has agreed to provide long-term hosting for all tools and content developed. Finally, a white paper will be released at the project's conclusion explaining what collaborators have learned about developing designing accessible digital humanities projects and making suggestions for "retrofitting" existing projects.

Building a Tool for the Analysis of Translations: The Case of Epistemic Modality in Edgar Allan Poe's Stories

Zupan, Simon
simon.zupan@uni-mb.si
University of Maribor, Slovenia

Juuso, Ilkka
ilkka.juuso@ee.oulu.fi
University of Oulu, Finland

Opas-Hänninen, Lisa Lena
lisa.lena.opas-hanninen@oulu.fi
University of Oulu, Finland

This paper investigates epistemic modality in Edgar Allan Poe's Gothic stories and their translations into Slovenian and Finnish. To facilitate this we have built a tool that allows the researcher to align the translations, search for the instances of epistemic stance and compare the original text with the translations. From the point of view of translation studies, epistemic modality is interesting in the context of Gothic stories, because it helps to create the mood of the story and any major shifts in the translation may cause a loss of the essence of the Gothic. On the other hand, the tool we have built is applicable to any investigation into a text and its translations and will, we hope, significantly aid those working in the field.

Epistemic modality has traditionally been regarded as the "manner in which the meaning of a clause is qualified so as to reflect the speaker's judgement of the likelihood of the proposition it expresses being true" (Quirk et. al. 1992: 219). In other words, it shows the position that speakers adopt with respect to the truth value of what they are saying (hence it is also referred to as epistemic stance). As Halliday (Halliday and Matthiesen 2004) has pointed out, speakers can adopt two extreme positions; they either qualify their proposition as true (e.g. "Translation scholars are in need of efficient electronic text-analysis tools.") or, alternatively, as not true ("Translation scholars are not in need of efficient electronic text-analysis tools."). These two positions are referred to as positive and negative polarity, respectively. More importantly, speakers can also adopt various positions in between the two poles. They can thus claim something to be "more" true ("Translation scholars are definitely in need of efficient text-analysis tools.") or "less" true ("Translation scholars are probably not in need of efficient text-analysis tools."). It is precisely these intermediate positions that are strictly referred to as epistemic modality because they indicate the speaker's (apparent) inability to ascertain the truth of the proposition they are making and consequently qualify it as polar.

This has important implications for the Gothic stories, in particular from the point of view of their translation into other languages. As Simpson (2004) has pointed out, authors of Gothic stories often employ epistemic modality to add uncertainty to them. These narratives thus abound in expressions such as "possibly", "perhaps", "undoubtedly", "it might have been", "I believe" and the like, which all indicate the protagonists' (apparent) inability to be able to tell what was behind a particular event or experience. In turn, this makes that same event or experience appear mysterious. It also lays ground for evoking in the reader the prototypical Gothic effects such as that of discomfort, uncanniness or eeriness. Consequently, these linguistic features of the original text need to be preserved in the translation. As our preliminary research has shown, however, this is not always the case. Comparison of one of the Slovenian translations with the original has shown that the translator failed to preserve epistemic modality in some of the examples by turning them into polar sentences. As a result, those passages in the target text lost some of their potential to evoke the same Gothic effects as their corresponding passages in the original.

In order to further investigate the extent of this phenomenon, we analyse several of Edgar Allan Poe's stories. These include *The Fall of the House of Usher, Berenice, The Masque of the Red Death, Metzengerstein*, and *Ligeia*, all of which can be seen as Gothic stories or stories with Gothic elements.

To facilitate the comparison of the source text and the target text(s), we have built ParallelTexts, a tool that allows the researcher to view two or three texts simultaneously in a synchronized manner. It makes use of XML markup, using pages and paragraphs, or other user-defined elements, for the purposes of synchronization of texts. There is support for marking up additions, deletions and other shifts in the text. The user interface shows a table-like view of the texts and allows for simultaneous scrolling of the texts. Figure 1 below shows the basic principle of the use interface, where the text in red on the left-hand side is highlighted to indicate that it is not found in the text on the right-hand side, ie the comparison version or the target text. Equally, the text in green on the right-hand side

indicates that this text has been added into the target text and has no equivalent in the original text.

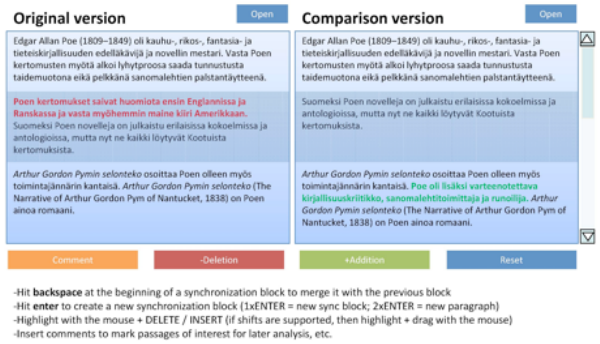


Fig. 1 A first version of ParallelTexts, with some basic instructions for using it

With appropriate input data, ParallelTexts allows us to examine various aspects of both the original text and its translation(s). First, it helps us identify most instances of epistemic modality in the original text and allows easy access to their corresponding sentences in the translation(s). Second, the tool allows us to statistically examine the texts and determine various quantitative features, among them the number of occurrences of epistemic modality and their distribution in both the original and the translation. Finally, the quantitative data collected with ParallelTexts will also allow us to qualitatively examine individual instances of epistemic modality in both texts and determine the cumulative effect of epistemic modality translation shifts on the translation as a whole. We also expect that the tool will prove to be useful for other similar translation studies text analyses. We believe that with some modification and appropriate input data, ParallelTexts could be used to compare any type of linguistic features in the original text and its translation. We wish to demonstrate the tool at DH2011 and all feedback and suggestions for further improvement of the tool will be most welcome.

References

Halliday, M.A.K. and Matthiessen (2004). *An Introduction to Functional Grammar.* London: Arnold.

Poe, E. A. (1982). *The Complete Tales and Poems of Edgar Allan Poe.* London: Penguin Books.

Quirk, R., Greenbaum, S., Leech, G. N., and Svartvik, J. (1992). *A Comprehensive Grammar of the English Language.* London and New York: Longman.

Simpson, P. (2004). *Stylistics: a resource book for students.* London and New York: Routledge.

Index of Authors