

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 2010

Conference Abstracts

King's College London, London
July 7 – 10, 2010



The 22nd Joint International Conference of the Association for Literary and Linguistic Computing and Association for Computers and the Humanities

and

The 3rd 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 Programme Committee

- Elisabeth Burr (ALLC)
- Richard Cunningham (SDH-SEMI)
- Jan-Christoph Meister (ALLC)
- Elli Mylonas (ACH)
- Brent Nelson (SDH-SEMI)
- John Nerbonne (ALLC and Chair)
- Bethany Nowviskie (ACH)
- Jan Rybicki (ALLC)
- John Walsh (ACH)

Local Organising Committee

- Sheila Anderson (Virtual conference)
- Tobias Blanke (Virtual conference)
- Gabriel Bodard (Workshop programme, and THATCamp)
- John Bradley (Workshop programme, and THATCamp)
- Sarah Davenport (Conference organiser and administrator)
- Mark Davies (Technical facilities)
- Camille Desenclos (Visiting fellow - editorial support)
- Hugh Denard (Performance strand)
- Stuart Dunn (Virtual conference)
- Mark Hedges (Virtual conference)
- Lorna Hughes (Co-chair; Virtual conference)
- Martyn Jessop (Publicity)
- John Lavagnino (Publicity)
- Michael Magruder (Performance strand)
- Willard McCarty (Busa Award lecture)
- Jamie Norrish (Conference website)
- Elena Pierazzo (Conference website; lead editor and technical support, Book of Abstracts)
- Torsten Reimer (Virtual conference)
- Harold Short (Co-chair)
- Helen Skundric (Conference organiser and administrator)
- Paul Spence (Co-chair)
- Simon Tanner (Sponsorship, and Exhibits)
- Charlotte Tupman (Editorial support, Book of Abstracts)
- Carrie van de Langenberg (Conference organiser and administrator)
- Paul Vetch (Co-chair)

ISBN 978-0-9565793-0-0

Published by

Office for Humanities Communication

Centre for Computing in the Humanities, King's College London

Conference logo and cover design by Damien Doherty

Texts encoded by Alberto Campagnolo, Camille Desenclos, Greta Franzini, John Levin and Nàdia Revenga Garcia.

© Copyright King's College London and the authors

Welcome from the Principal, King's College London

Prof. Trainor, Rick (MA DPhil FRHistS AcSS FKC)

King's College London is very proud of its long tradition in the application of advanced computational methods in arts and humanities research, as well as the significant achievements of your conference hosts, the Centre for Computing in the Humanities and the Centre for e-Research.

King's has long seen humanities computing as strategically significant, as reflected by its specific mention in the current and previous two instances of the College's Strategic Plan. Since the School of Arts and Humanities has been a traditional strength of the College, it is particularly heartening, as well as entirely appropriate, that the digital humanities should have become and remains one of the College's and the School's distinctive intellectual features.

King's College London is delighted to host the Digital Humanities 2010 conference, and it gives me great pleasure to welcome you to London, the College and the conference. I wish the conference every success.

Welcome from the Head of the School of Arts and Humanities

Prof. Palmowski, Jan (DPhil)

King's College London

The School of Arts and Humanities has benefited significantly from the development of the Centre for Computing in the Humanities, which it has nurtured, building on foundations laid in the early 1970s. It is proud of the prominent role played by CCH and its staff in the national and international digital humanities, and of having the field's first academic department and its first PhD programme. We have also gained great benefit from the wider institutional and national roles of the Centre for e-Research - and its national predecessor, the Arts and Humanities Data Service, whose closure is much regretted.

The digital humanities continues to play a key role in the life of the School and is seen as strategically important as the School moves forward, not only in terms of an increasing range of collaborative research and teaching, but also within key thematic areas of development, including digital and visual culture.

On behalf of the School of Arts and Humanities, I bid you a warm welcome to the College and to the DH2010 conference.

Chair of International Programme Committee

Nerbonne, John

University of Groningen

If you ever get the chance to join a program committee for a Digital Humanities (DH) conference, you should grab it! It is exciting, gratifying and energizing to work on the program committee (PC), where one sees – even more than at the conference itself – how the field is growing, seeking new directions, reviewing past successes and failures, and sifting through the earlier proposals, techniques and analyses that constitute digital humanities. There are new names and affiliations each year, a sure sign of the field's vitality, and there are lots of great new ideas being described for the first time!

The PC tried to break new ground in asking a member *in* (and not merely *near*) the DH community to share her vision in a plenary address, feeling the field has matured enough to benefit from that level of reflection, and we tried to increase the variety of speakers by screening the early scores for double submissions by the same author. We suggested in such case that one paper might better be presented as poster. Of course, this was also prompted by the large number of submissions we received.

Growth is a great sign that we're on the right track, but it also presents a problem we need to solve. There were three times as many abstracts submitted as there were time slots for papers, and the competition for panels was even stiffer. This year's PC sent letters of rejection in which we tried to use these numbers to console contributors whose abstracts hadn't been accepted, but some took our citing the numbers the wrong way – as a sign that we'd been pleased at rejecting so many abstracts. Of course, we *were* pleased at the high level of interest, but we would have been more pleased to include more of the submissions somewhere in the program.

The PC was a privilege to work with – not only professional and conscientious in considering the submissions and planning the program, but also in reviewing a set of suggestions on how things might be improved in the future! The decisions about future conferences won't be up to this year's PC, but cutting back on the number of panels, stimulating interest in poster sessions, and exploiting the opportunities that satellite workshops offer are three of the options we suggested for further consideration.

Many hands may light work, and not only in manual tasks. I especially wish to thank the vice chair, Bethany Nowviskie of Virginia, who was unstinting in her energies for the good cause and judicious in her advice. Thanks are also due to Sara Schmidt at the University of Illinois and the local administration group at King's College, London, all of whom have done a tremendous amount of work in preparing the conference program. Finally, the local organizers, especially Harold Short, Paul Spence, Paul Vetch and Elena Pierazzo, were cooperative in every way when it came to finding rooms, working out a schedule, and all the myriad other things that you fortunately needn't know of.

Introduction & Welcome from the Local Hosts

1. Welcome!

Welcome to Digital Humanities 2010 (*DH2010*) from the local organisers: the Centre for Computing in the Humanities and the Centre for e-Research at King's College London. It will be hard to match the vibrancy of *DH2009* in Maryland, but this has been our goal, and from the range and quality of the papers, panels and posters, the intellectual excitement we hope will be engendered by the Performance strand, as well as the numbers of people who have registered to attend the conference, we can hope that this conference will be memorable in its own right.

It is a particular pleasure to welcome the large number of early-career researchers and students who will be presenting at and attending the conference. This has been made possible in significant part by the generosity of funders in providing bursaries: the European Science Foundation; the Daiwa Anglo-Japanese Foundation; and the Alliance of Digital Humanities Organisations. In addition the local organisers have been able to offer Student Assistant bursaries. (See further details below.)

The presenters of papers and posters are drawn from over 20 countries, and as this volume went to press, registrations had been received from five continents and over 25 countries, making for a truly international event. We hope and believe this will be further enhanced by the international multi-cultural character of the host city – London.

London is one of the world's major capital cities, and one of the most cosmopolitan. Its cultural diversity affects many aspects of the city's life, including cuisine, with what must now be the most ethnically diverse range of food in any capital city. It must also be one of the easiest, and perhaps cheapest, places in the world to get to, and is an exceptionally easy city to move around in. It has some of the world's most important cultural institutions, and is a major European and international venue for many forms of the creative arts, including theatre and music. We hope you will take the opportunity while in London to explore and experience some the city's rich cultural diversity.

It is worth noting that one the major cultural institutions referred to, the British Library, will have a stand in the Great Hall throughout the conference, providing information among many other things about the major exhibition it will be hosting from Autumn 2010 on *Digital Scholarship*.

2. The *Digital Humanities* Conferences

DH2010 is the fifth conference to have the designation 'Digital Humanities', the first having been in Paris in 2006, following the establishment of the Alliance of Digital Humanities Organisations (ADHO), the conference's main sponsor.

Its antecedents go back much further, however, to the conferences organised individually by ADHO's constituent organisations: the Association for Literary and Linguistic Computing (ALLC) – whose first conference was held in 1974; the Association for Computers and the Humanities (ACH) – 1981, and the Society for Digital Humanities / Société pour l'Étude des Médias Interactifs (SDH-SEMI), a Canadian national association with continuing annual national conferences.

In the late 1980s ALLC and ACH established a joint international conference, the first taking place at the University of Toronto in 1989. Since that time the conference venues have alternated between North America and Europe. ADHO has to date continued this tradition, so *DH2009* was held at the University of Maryland and *DH2011* will be hosted at Stanford University in California. As scholarly activity in the digital humanities continues to develop around the world, this pattern may need to change.

3. Alliance of Digital Humanities Organisations

ADHO was established to provide a framework for international collaboration in digital humanities developments and activities. At the beginning the key areas of focus were the annual conference and the

range of publication venues, which now include not only the print & online journal *LLC: The journal of digital scholarship in the humanities*, published by Oxford University Press, but also two peer reviewed electronic journals – *Digital Humanities Quarterly* and *Digital Studies/Le Champs Numérique*, as well as the online seminar Humanist – now in its 24th year, still edited by Willard McCarty – and two monograph series, one published by Ashgate and the other by Illinois University Press.

At present ADHO's constituent organisations are ALLC, ACH and SDH-SEMI. ALLC and ACH established ADHO as a collaborative venture in 2005, with SDH-SEMI admitted to membership in 2006. From the outset a key objective has been to foster the development of other regional and national digital humanities associations, with a view to expanding the institutional membership of ADHO.

ADHO is an alliance of organisations, and the current constituent organisations are individual membership associations, where individuals become members by subscribing to LLC. This model is intended to continue as new regional associations are formed, with as much support from ADHO as possible and appropriate. A recent development of a different kind is worthy of mention, however. This is CenterNet, an international association of centres and departments of digital humanities, which has been developing rapidly under the leadership of Neil Fraistat, Kay Walter and a number of colleagues, and which now has over 100 members. Discussions are in progress to work out how best to accommodate CenterNet within the ADHO umbrella, something keenly desired by all concerned.

4. Bursaries and young scholars

It is particularly gratifying that *DH2010* promises to have a significant number of younger scholars, including doctoral students and early-career researchers. In part this may be attributed to a number of funding schemes – bursaries of various kinds – that support their attendance, and in part to the vibrancy of the field, which appears to excite the intellectual passions of increasing numbers of young scholars.

Bursaries for *DH2010* have come from four sources. For each Digital Humanities conference ADHO offers ten bursaries to support young scholars, and all ten were awarded for *DH2010*. We are also grateful to the European Science Foundation, whose bursary funding has provided financial assistance to four young European scholars. ADHO has graciously agreed to fund the attendance of not only the ADHO but also the ESF award winners at the Conference Dinner.

We are similarly grateful to the Daiwa Anglo-Japanese Foundation which has awarded travel grants to make it possible for five young Japanese researchers to attend the conference. Each recipient is receiving corresponding support for accommodation and living expenses from their sponsoring institution in Japan, and the local organisers have waived their registration fees and have offered free places at the Conference Dinner.

The local organisers this year introduced Student Assistant bursaries. These provide free conference registration, accommodation costs (for students from outside London) and living expenses plus free attendance at the Conference Dinner for students who are willing to undertake four hours work per day to support the running of the conference. The rationale of the scheme is that the conference will benefit from the commitment and enthusiasm of the Student Assistants, and the young people will benefit both intellectually and social, not only from being able to attend a conference that might otherwise have been impossible, but also from being able to meet and interact with a wide circle of digital humanities scholars.

5. Conference programme

For the conference's wide-ranging **academic programme**, we owe a considerable debt of appreciation and thanks to the International Programme Committee for *DH2010*, chaired by John Nerbonne. The range of topics covered in the papers, panels and posters continues to expand. Long-standing areas of intellectual enquiry such as authorship attribution, literary and linguistic analysis, text encoding and data modelling continue to have strong representation, but so too do much 'newer' interests including GIS & mapping and social networking tools. Our plenary speakers promise to provide rich stimulation, from Chuck Henry in the opening session to the closing keynote by one of the field's most promising younger scholars, Melissa Terras.

In between we will have the Busa Lecture. The **Busa Prize** is an award by ADHO for lifetime's achievement in the digital humanities. It is presented every three years, and at the conference where the

award is presented, the prize-winner is invited to give a plenary lecture. We are fortunate that 2010 is a Busa year, and the award and lecture will take place on Thursday 8 July. The award winner is Professor Emeritus Joseph Raben, one of the pioneers of computational methods in humanities research.

6. The Virtual Conference

There will be many opportunities for virtual participation in the conference. A focal point for access and dissemination will be <http://www.arts-humanities.net>, the knowledge base and community forum for the digital arts and humanities, with links from the conference website. Arts-humanities.net is based at the Centre for e-Research, one of the hosts of *DH2010*.

These include:

Twitter: the hash-tag #dh2010 has been in use for many months, and it is expected that this will become particularly active during the conference, following the pattern established in *DH2009* at the University of Maryland.

Podcasts: all three plenary sessions will be podcast, and if possible streamed live, via arts-humanities.net. Other podcasts will be made available each day of the conference. It is hoped that it will be possible to provide a near-complete set of podcasts covering all conference sessions.

Blog(s): at least one conference blog will be available on arts-humanities.net.

Student Assistants and audio-visual mashups: a number of students have been signed up to assist with the running of the conference. Each day, some of these students will be asked to record conference sessions, while others will record brief on-the-spot interviews with speakers and delegates. Some of these recordings will be archived and made available as podcasts, while a number will be taken by a small team of ‘editors’ in order to create what we hope will be interesting audio-visual mashups that capture the spirit and some of the highlights of the conference. New mashups will be produced each day, for screening in the **Anatomy Theatre and Museum** (ATM) as well as being made available via *arts-humanities.net*.

Posters: student assistants will be assigned the task of recording brief interviews with each poster presenter, and the complete set of recordings will be made available on *arts-humanities.net*.

Slides: Each presenter will be invited to provide a copy of any set of slides they use, for inclusion on a *DH2010* site on slideshare.net. In addition, each poster display will be photographed and published – with the presenters’ permission – as a set of slides on the same site.

7. Digital Scholarship at King’s College London

Running alongside the conference and the virtual conference, there will be an informal series of events in the ATM presenting various aspects of digital scholarship across the disciplines at the host institution, King’s College London. Presentations, workshops and demonstrations will run throughout the week and conference delegates are welcome to attend and participate in these events. There will be a combination of presentations, discussions and workshops (on Monday, Tuesday and Wednesday morning) and drop-in demonstrations and opportunities for hands-on experimentation (Thursday and Friday). These will focus on four key themes: **research infrastructures; tools and methods for advanced research; advanced techniques in practice; and interdisciplinary collaboration**. There will also be an opportunity to see some student work in the digital arts and humanities, and to find out more about digital humanities MA programmes offered at King’s College London, including the MA in Digital Humanities, MA in Digital Culture & Technology, and the new MA in Digital Asset Management, due to start in October 2010.

8. Performances

The conference theme is ‘Cultural expression, old and new’, and in support of this, *DH2010* has a significant **Performance** strand running through it. This begins in the opening session on Wednesday 7 July with illustrated dialogues on ‘performance and research’ and an opening keynote address by Chuck Henry on the preservation and interpretation of performance. It also includes a performance element at the reception in the Conservatory at the Barbican Centre on Friday 9 July involving students

from Guildhall School of Music and Drama. Throughout the conference, two art installations will be featured in the Great Hall: a participative artwork by Ele Carpenter entitled *Embroidered Digital Commons*, developed as part of her *Open Source Embroidery* project; and two digitally-based art installations by Michael Takeo Magruder: Vanishing Point(s) and Communion. In addition there will be more ‘traditional’ performances to accompany the opening and closing receptions, and following the conference dinner.

9. Social Programme

The **social programme** revolves around three receptions and the Conference Dinner. The first reception will follow the opening session on Wed 7 July, and second will take place in the remarkable Conservatory at the Barbican Centre, accompanied by a performance as described above. The Conference Dinner will close the conference on Sat 10 July. It will be held in the Great Hall of Lincoln’s Inn, a short walk from the main conference venue on the Strand, and will be preceded by a reception on the Terrace, looking out over the gardens. There are even a couple of informal guided walks following the first two receptions.

For good measure, a choice of three excursions is available on Sunday 11 July, very different in character from each other: a full day at Hampton Court, including its world-famous flower show as an additional extra; the Tate-to-Tate experience which includes guided tours of Tate Britain and Tate Modern, with a boat trip down the Thames in between; and a guided tour of Shakespeare’s Globe, including all-day access to its exhibition.

We wish you welcome to London, to King’s College and to the *DH2010* conference, and trust that you will find rich intellectual and social enjoyment during your stay.

Sheila Anderson, Lorna Hughes, Harold Short, Paul Spence, Paul Vetch

Conference Hosts

1. Centre for Computing in the Humanities

The application of computational methods in arts and humanities research has a long history at King's, going back to the early 1970s. The first applications here, as in many other places at the time, were in developing electronic concordances of literary works. Much of the initial impetus was provided by one of the European pioneers of humanities computing, Professor Roy Wisbey, who came to King's as Professor of German in 1971 from Cambridge, where he had already established the Centre for Literary and Linguistic Computing. The other early UK centre of humanities computing was at the University of Oxford. The founding meeting of the ALLC took place at King's in 1973, with Wisbey elected as its first Chair.

In 1987, following the merger of King's with two other University of London colleges, the computing services of the 'new' King's College London were structured such that one group of staff had the support of humanities departments as a specified area of responsibility. An early task for this group was to develop, in collaboration with the Faculty of Arts and Music, an undergraduate minor programme in Humanities Computing, which was initiated in 1989. At the same time, collaborative research with humanities scholars began to develop, to the extent that a Research Unit in Humanities Computing was formed in 1992 as a joint venture between the Faculty and Computing Services. The success of this unit led in turn to its expansion in 1995 into the Centre for Computing in the Humanities (CCH), on the same joint funding basis.

The basis of both the teaching and research carried out by CCH staff was fundamentally academic – as well as collaborative – in character, and the College and the School of Humanities (as the old Faculty had now become) decided in 2002 to move CCH fully into the School as an academic department, the first academic department of its kind in the world. At the time we decided not to change the name to 'Department' for 'brand recognition' reasons. It has recently been decided, however, that a change of name is now desirable, and from the start of the 2010-11 academic year, CCH will become the Department of Digital Humanities (DDH).

CCH has continued to develop its academic profile. It is now responsible for two Masters programmes - *Digital Humanities and Digital Culture & Technology*. A new MA in *Digital Asset Management* will begin in Autumn 2010 as a joint development by CCH and the Centre for e-Research (CeRch). In 2002 CCH introduced the world's first PhD programme in Digital Humanities, and currently has 12 registered students, all but three jointly supervised with departments in Humanities, the Social Sciences and Computer Science.

At any one time CCH is involved in 30 or more major research projects. All are collaborative with partners from across the Humanities – and some Social Science – disciplines, with CCH staff having primary responsibility for technical research. With its partners it has generated over 17 million GBP in research income since 2000. This level of success stems in part from the collaborative nature of its research engagements, in part from the research excellence of the School of Humanities, and in part from the critical mass of expertise CCH has developed across a wide range of discipline areas and technologies relevant to the digital humanities. The discipline areas include classical studies, medieval studies, history – including prosopography, literature, language, music, theatre and performance. The technologies include text modelling and encoding, digital publications, text analysis, database analysis and design, visual and interface design and development, visualisation including mapping and GIS, and 3D visualisation and virtual worlds research. A small sample of completed and current projects includes: Prosopography of Anglo-Saxon England, Fine Rolls of Henry III, Clergy of the Church of England Database, Inscriptions of Aphrodisias, Inscriptions of Roman Tripolitania, Chopin First Editions Online, Centre for the History and Analysis of Recorded Music, Corpus Vitrearum Medii Aevi, and Jane Austen's Fictional Manuscripts. Details of these and other CCH projects and links to their websites may be found on the CCH website at <http://www.kcl.ac.uk/research/projects>.

Integrated with CCH's core activities are two world-leading groups which work extensively with cultural, heritage and creative partners:

King's Visualisation Lab has a leading international reputation in 3D visualisation and virtual worlds research. KVL delivers 'an authentic, beautiful experience' through their visualisations which, for example, extend from the Theatre of Pompey the Great in Rome through the Roman frescos at Boscoreale to 'How Kew Grew', which shows the development of Kew Gardens from the building of Kew Palace in 1631 to the present day. KVL has been a prime mover in the development of the London Charter, which is of growing international importance in defining standards in the digitisation of cultural heritage, already having been translated from English into a number of other languages, and formally adopted by the cultural heritage authorities in several countries.

King's Digital Consultancy Services delivers digital change management and strategic consultancy to the cultural and commercial sectors worldwide. KDCS founded the Digital Futures Academy run annually in London and Sydney, Australia. Examples of KDCS activity include: digitisation of the Dead Sea Scrolls; feasibility and business planning for the National Library's of Ireland, Scotland and Wales; plus development of digital strategy for National Museums Northern Ireland.

As an academic department, CCH submitted a return in the UK's national Research Assessment Exercise (RAE) 2008, including six members of the Centre for e-Research. The joint submission received an exceptionally high rating, with 35% of its research output placed in the highest category ('world-leading') – the highest percentage in its sector – and 30% in the second category ('internationally excellent').

One of the most ambitious projects in which CCH has been involved was the AHRC ICT Methods Network (2005-2008), a core component of the ARHC programme 'ICT in Arts and Humanities Research', led by David Robey, then Professor of Italian at the University of Reading (and President of ALLC). The Methods Network was based at King's, with Lorna Hughes as Programme Manager and Marilyn Deegan and Harold Short as Co-Directors. The aim of the Methods Network was to explore and promote the use of advanced ICT methods in arts and humanities research. Its programme included two major strands: a series of 'expert seminars', which brought together invited specialists to describe current cutting edge work in their discipline areas; and a number of workshops and other self-directed activities, each receiving a small amount of financial and administrative support.

The expert seminars were the basis of the Digital Research in the Arts and Humanities print series currently in publication by Ashgate (7 of the 9 commissioned volumes are in print, with the others to follow by early 2011. (The series is on display during the conference at the Ashgate stand in the Great Hall.) The workshops succeeded in engaging over 1,000 UK participants plus a number from Europe and North America.

The development of CCH has owed a great deal to the shared vision of senior College administrators and senior Arts and Humanities academics. These include: Roy Wisbey (mentioned above); Professor Barry Ife in his roles as Head of School of Humanities, Vice-Principal and Acting Principal; the current Principal, Professor Rick Trainor – himself an historian and digital humanist; and a succession of Heads of School, including David Ricks, Ann Thompson and the current Head, Professor Jan Palmowski.

Professor Ife moved from King's to become Principal of Guildhall School of Music and Drama, which is a very active partner in the Performance strand of *DH2010*.

2. Centre for e-Research

Launched in April 2008 following the sad demise of the Arts and Humanities Data Service (AHDS) and the end of the AHRC ICT Methods Network project, the Centre for e-Research (CeRch) is a research centre located in Information Services and Systems (ISS). It is an academic Centre outside the School and Department structure aimed at facilitating interdisciplinary, institutional, national and international collaboration. The Centre's strengths are in sustainable e-infrastructures for research; digital libraries and digital archives including data use, creation, curation and preservation; researcher practices in the digital domain; and ICT-Methods with particular expertise in e-Science, geo-spatial and geo-temporal methods, text mining, textual analysis, and use of grids. The Centre is unusual in that it is both an academic centre researching, publishing and teaching in its areas of expertise – including contributing to the UK's 2008 Research Assessment Exercise (RAE) in a Unit of Assessment with the Centre for

Computing in the Humanities (CCH) and, again with CCH, developing a new MA programme in Digital Asset Management – and a focus for ISS-related activities supporting e-research, data management, and the curation and preservation of research data.

This unusual profile reflects the origins of the Centre arising, as it did, from the ashes of the AHDS Executive which was hosted at King's (located in ISS) from its inception in 1996 until its national funding was withdrawn in March 2008. Following a feasibility study by Harold Short and Lou Burnard, the AHDS was established as a distributed organisation with six centres: The Executive at King's, Literature, Languages and Linguistics at Oxford, Archaeology at York, History at Essex, Performing Arts at Glasgow, and Visual Arts at University College for the Creative Arts, Farnham. Over its twelve year history, the AHDS built up a world class reputation in the management of digital content in the arts and humanities, including: standards and guides to best practice for data creation, use, and preservation; tools and systems for creating, managing, preserving data, and providing access to data; underpinned by a range of strategies and policies for collections development, appraisal, management and preservation. The AHDS was also an early entrant into exploring and researching e-Science methods and tools for the arts and humanities (Cyberinfrastructure in North America and elsewhere). It is this legacy that, with the support of colleagues at King's and elsewhere, the Centre for e-Research has taken forward over the last two years.

The Centre comprises a mix of academic researchers, librarians, systems architects and analysts, developers and programmers, and departmental and project managers and administrators. The Centre works collaboratively with researchers, research teams and groups, and as partners in research projects across King's College London. It also works in partnership with other institutions in the UK and mainland Europe, and internationally with HE library and research institutes. At the time of writing it has a project portfolio of 18 projects funded by the Joint Information Systems Committee (JISC), the Arts and Humanities Research Council (AHRC), the Engineering and Physical Sciences Research Council (EPSRC), and the European Commission.

It remains the host of two key projects started under the auspices of the AHDS and the Methods Network: the Arts and Humanities e-Science Centre (AHeSSC), and *arts-humanities.net*. AHeSSC forms a critical part of the AHRC-EPSRC-JISC initiative on e-Science in Arts and Humanities research. It supports, co-ordinates, promotes, and researches into e-Science across arts and humanities disciplines, and liaises with the e-Science and e-Social Science communities, in particular computing and information sciences. AHeSSC is helping to understand the impact of advanced use of ICT methods and tools encompassed in e-Science, the value and the challenges that it brings to research practices and knowledge generation, and how we use these insights to construct research infrastructures that support and enhance epistemic practices in the humanities and arts.

Arts-humanities.net aims to support and advance the use and understanding of digital tools and methods for research and teaching in the arts and humanities by providing:

- Information on projects creating and using digital content, tools and methods to answer research questions
- Information on tools and methods for creating and using digital resources
- A listing of expert centres and individual researchers
- A library documenting lessons learned through case studies, briefing papers, and a bibliography

It is the product of the CCH AHRC ICT Methods Network (2005-2008) and the AHDS AHRC/JISC funded ICTGuides project. *Arts-humanities.net* is a community resource and we invite you to join as a member at <http://www.arts-humanities.net/>. Members are encouraged to contribute information about their own projects, tools and research, to publicise events, conferences, and job vacancies, and to take part in and set up discussion forums.

CeRch is taking an active role on the European stage playing a leading role in two large infrastructure projects. The Digital Research Infrastructure for the Arts and Humanities (DARIAH) is a large-scale preparatory project to identify and research the key components necessary for an pan-European infrastructure that supports and enhances scholarship in the arts and humanities. The mission of DARIAH is to enhance and support digitally-enabled research across the humanities and arts. It aims to develop and maintain an infrastructure in support of ICT-based research practices and will work with communities of practice to:

- 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
- Improve research opportunities and outcomes through linking distributed digital source materials of many kinds
- Exchange knowledge, expertise, methodologies and practices across domains and disciplines.

CeRch is leading on the project's strategic and technical work packages.

In the year the world commemorates the 65th anniversary of the liberation of Auschwitz, the European Commission is to fund a European Holocaust Research Infrastructure (EHRI) project to support research into the Holocaust. EHRI's main objective is to support the European Holocaust research community and help initiate new levels of collaborative research through the development of innovative methodologies, research guides and user-driven transnational access to research infrastructures and services. EHRI will design and implement a Virtual Research Environment (VRE) offering online access to a wide variety of disparate and dispersed key Holocaust archival materials and to a number of online tools to work with them. EHRI sets out to transform the data available for Holocaust research around Europe and elsewhere into a cohesive corpus of resources. CeRch is leading on the development of the VRE and on research into scholarly practices that will inform the structuring, modeling and integration of the digital content, and the functions of the VRE.

CeRch is also pursuing research into the potential of linked data to support scholarship and create a 'web of knowledge'; citizen cyberscience and crowdsourcing to enable collective content development and distributed thinking; and 'collective intelligence' approaches both to create a sense of community and to capture expertise and knowledge.

None of this would have been possible without the sanction and support of the Principal, Professor Rick Trainor, the Chief Information Officer and College Librarian, Karen Stanton, the VP for Arts and Science, Professor Keith Hoggart, and Dr Trudi Darby, Deputy Head of Administration (Arts & Sciences) all of whom supported the establishment of the Centre following the unexpected demise of the AHDS, and whose support since then has enabled CeRch to develop into a thriving research centre.

3. King's College London

King's College London is one of the top 25 universities in the world (Times Higher Education 2008) and the fourth oldest in England. A research-led university based in the heart of London, King's has more than 21,000 students from nearly 140 countries, and more than 5,700 employees. In the UK's 2008 Research Assessment Exercise (RAE), 23 departments were ranked in the top quartile of British universities, and the College is in the top seven UK universities for research earnings with an overall annual income of nearly £450 million.

The College has nine Schools of Study, and a particularly distinguished reputation in the humanities, law, the sciences (including a wide range of health areas such as psychiatry, medicine and dentistry) and social sciences including international affairs. It has played a major role in many of the advances that have shaped modern life, such as the discovery of the structure of DNA and research that led to the development of radio, television, mobile phones and radar.

Famous names associated with King's include Rosalind Franklin and Maurice Wilkins of DNA fame, James Clerk Maxwell, Florence Nightingale, James Black and Desmond Tutu. Among many well-known writers associated with King's are John Keats, Charles Kingsley, John Ruskin, Sir William Gilbert, Thomas Hardy, Somerset Maugham, Anita Brookner, Sir Arthur C Clarke, Radclyffe Hall and Hanif Kureishi.

4. School of Arts & Humanities

The 2008 RAE confirmed the world-class standard of research undertaken in the School of Arts & Humanities and its leading international reputation. Three of 14 departments are distinguished by the highest proportion of 'world-class' research undertaken in any UK university in their fields. When measuring work rated as 'world class' and 'internationally excellent' (4* and 3*), 13 out of 14 departments were ranked amongst the top six departments nationwide.

In addition to its many traditional strengths, over recent years the School has developed a number of new areas of academic activity, all of which have been particularly successful. Film Studies, American Studies, and the Centre for Computing in the Humanities have all been recent creations and were submitted for the first time to the 2008 RAE. All were ranked in the top five nationally. From Summer 2010 another recent initiative - the Centre for Culture, Media and the Creative Industries – will be grouped with CCH, the Department of Music and the Department of Film Studies in a ‘Creative Arts’ thematic cluster with a view to fostering further collaborative work within the cluster and across the School.

All departments in the School are engaged in the digital humanities – in research, and increasingly in teaching. This represents a wide range of discipline areas: Byzantine & Modern Greek Studies; Classics; English; European Studies; Film Studies; French; German; History; Music; Philosophy; Portuguese & Brazilian Studies; Spanish & Spanish American Studies; Theology & Religious Studies. Engagements are at every level, from PhD students and junior lecturers to Professors and Heads of Department.

Acknowledgments

We are very pleased to acknowledge support and assistance from many quarters which has made possible the planning and preparations for the DH2010 Conference. These include:

The host institution, King's College London, starting with the Principal, Professor Rick Trainor, and covering a large number of departments and individuals, including Dr Trudi Darby, Deputy Head of Administration (Arts & Sciences).

Professor Jan Palmowski, Head of the School of Arts and Humanities, along with many of our colleagues in the School, in particular Christine Saunders and Wendy Pank in the School Office, and research partners across the School.

The Chief Information Officer and College Librarian, Karen Stanton and many of her staff in Information Services and Systems, in particular Lucy Burrow, Head of IT Policy and Process.

The Conference's sponsors and exhibitors: the European Science Foundation, the Daiwa Anglo-Japanese Foundation, Oxford University Press; Ashgate Publishing, the British Library, Maney Publishing, MIT Press, Vishal Information Technologies Ltd; and the Alliance of Digital Humanities Organisations.

The Guildhall School of Music and Drama for their engagement in the Performance strand of the conference, in particular the Principal, Professor Barry Ife, the Assistant Principal (Research and Academic Development), Helena Gaunt, and Pamela Lidiard, Deputy Head of Keyboard Studies.

The *ConfTool Pro* conference software developed by Harald Weinreich, which has now been used for a number of *Digital Humanities* conferences. Also Sara Schmidt of the University of Illinois Urbana-Champaign, for substantial assistance with *ConfTool*.

The International Programme Committee.

The Local Organising Committee.

Bursary Winners

1. Alliance of Digital Humanities Organisations

- Hui, Barbara (Comparative Literature, UCLA)
- Finn, Ed (Dept of English, Stanford University)
- Bunde, Janet (University Archives, New York University)
- Zhu, Jichem (Digital Media, University of Central Florida)
- Sookhanaphibarn, Kingkarn (Kinugasa Research Organization, Ritsumeikan University)
- Fu, Liu (Old Dominion University)
- Sokól, Małgorzata (Department of English, Szczecin University)
- Büchler, Marco (Computer Science, Leipzig University)
- Sainte-Marie, Maxime B. (Cognitive Computer Science, Université du Québec à Montréal)
- Organisciak, Peter (University of Alberta)

2. European Science Foundation

- Guy, Georgina (Department of English, King's College London)
- Howell, Sonia (An Foras Feasa, National University of Ireland, Maynooth)
- Wiersma, Wybo (Centre for Computing in the Humanities, King's College London)
- Zöllner-Weber, Amélie (Uni Digital, University of Bergen)

3. Daiwa Anglo-Japanese Foundation

- Iwasaki, Yoichi (Indian Philosophy & Buddhist Studies, University of Tokyo)
- Kobayashi, Yuichiro (Graduate School of Language & Culture, Osaka University)
- Matsuda, Kuninori (International Institute for Digital Humanities, Tokyo)
- Takahashi, Koichi (Indian Philosophy & Buddhist Studies, University of Tokyo)
- Ota, Asuka (Graduate School of Library, Information, and Media Studies, University of Tsukuba)

4. DH2010 Student Assistant Bursaries

- Mikulec, Anna (Pedagogical University of Krakow)
- Bajak, Barbara (Pedagogical University of Krakow)
- Nocera, Claudia (University College London)
- Csorba, Gabriel (ELTE University, Budapest)
- Civiliene, Gabriele (King's College London)
- Fanzini, Greta (King's College London)
- Doutsou, Ioanno (King's College London)
- Remy, Jana (University of California, Irvine)
- Lo, Jennifer (King's College London)

- Levin, John (King's College London)
- Szosta, Katarzyna (Pedagogical University of Krakow)
- Bolick, Laura (Open University)
- Schuech, Lena (University of Hamburg, Institute for Modern German Literature)
- Gruenhage, Lisa (University of Hamburg)
- Borek, Luise (Trier University)
- Lame, Marion (Université de Provence (France) / Università di Bologna (Italy))
- Buning, Marius (European University Institute)
- Romanello, Matteo (King's College London)
- Revenga Garcia, Nàdia (King's College London)
- Anderson, Stephen Cosmo (University College London)
- Salyers, Tom (King's College London)

Additional Student Assistants

- Campagnolo, Alberto (King's College London)
- Desenclos, Camille (King's College London)
- Díaz Bravo, Rocío (Queen Mary)

Table of Contents

List of Reviewers.....	1
Plenary Sessions	
To Hold Up a Mirror: Preservation and Interpretation of Performance in a Digital Age	
<i>Henry, Charles J.</i>	7
Humanities Computing in an Age of Social Change	
<i>Raben, Joe</i>	8
Present, Not Voting: Digital Humanities in the Panopticon	
<i>Terras, Melissa</i>	9
Art Installations	
Vanishing Point(s) and Communion	
<i>Magruder, Michael Takeo; Denard, Hugh</i>	13
The Embroidered Digital Commons: Rescension	
<i>Carpenter, Ele</i>	17
Pre-conference Workshops	
Access to the Grid: Interfacing the Humanities with Grid Technologies	
<i>Dunn, Stuart</i>	21
Content, Compliance, Collaboration and Complexity: Creating and Sustaining Information	
<i>Evans, Joanne; Henningham, Nikki; Morgan, Helen</i>	22
Text Mining in the Digital Humanities	
<i>Heyer, Gerhard; Büchler, Marco; Eckart, Thomas; Schubert, Charlotte</i>	23
Introduction to Text Analysis Using JiTR and Voyeur	
<i>Sinclair, Stéfan; Rockwell, Geoffrey</i>	25
Designing a Digital Humanities Lab	
<i>Veomett, Angela</i>	26
Peer Reviewing Digital Archives: the NINES model	
<i>Wheeler, Dana; Mandell, Laura</i>	27
Panels	
Digital Literacy for the Dumbest Generation - Digital Humanities Programs 2010	
<i>Clement, Tanya; Jannidis, Fotis; McCarty, Willard</i>	31
Computational approaches to textual variation in medieval literature	
<i>van Dalen-Oskam, Karina; Thaisen, Jacob; Kestemont, Mike</i>	37
Building the Humanities Lab: Scholarly Practices in Virtual Research Environments	
<i>van den Heuvel, Charles; Antonijevic, Smiljana; Blanke, Tobias; Bodenhamer, David; Jannidis, Fotis; Nowviskie, Bethany; Rockwell, Geoffrey; van Zundert, Joris</i>	44
Wargames in a Digital Age	
<i>Kirschenbaum, Matthew; Juola, Patrick; Sabin, Philip</i>	46

Scanning Between the Lines: The Search for the Semantic Story <i>Lawrence, K. Faith; Battino, Paolo; Rissen, Paul; Jewell, Michael O.; Lancioni, Tarcisio.....</i>	52
Standards, Specifications, and Paradigms for Customized Video Playback <i>McDonald, Jarom Lyle; Melby, Alan K.; Hendricks, Harold.....</i>	61
The Origins and Current State of Digitization of Humanities in Japan <i>Muller, A. Charles; Hachimura, Kōzaburō; Hara, Shoichiro; Ogiso, Toshinobu; Aida, Mitsuru; Yasuoka, Koichi; Akama, Ryo; Shimoda, Masahiro; Tabata, Tomoji; Nagasaki, Kiyonori.....</i>	68
Born Digital: The 21st Century Archive in Practice and Theory <i>Redwine, Gabriela; Kirschenbaum, Matthew; Olson, Michael; Farr, Erika.....</i>	71
Networks of Stories, Structures and Digital Humanities <i>Salah, Almila Akdag; Nooy, Wouter De; Borovsky, Zoe.....</i>	76
Understanding the 'Capacity' of the Digital Humanities: The Canadian Experience, Generalised <i>Siemens, Ray; Eberle-Sinatra, Michael; Siemens, Lynne; Sinclair, Stéfan; Brown, Susan; Timney, Meagan; Rockwell, Geoffrey.....</i>	82
Coalition of Humanities and Arts Infrastructures and Networks - CHAIN <i>Wynne, Martin; Anderson, Sheila; Fraistat, Neil; Kainz, Chad; Krauwer, Steven; Robey, David; Short, Harold.....</i>	84
Papers	
Character Encoding and Digital Humanities in 2010 – An Insider's View <i>Anderson, Deborah.....</i>	87
Semantic Cartography: Using RDF/OWL to Build Adaptable Tools for Text Exploration <i>Ashton, Andrew.....</i>	89
Using Wikipedia to Enable Entity Retrieval and Visualization Concerning the Intellectual/Cultural Heritage <i>Athenikos, Sofia J.....</i>	92
Mapping the World of an Ancient Greek Historian: The HESTIA Project <i>Barker, Elton; Pelling, Chris; Bouzarovski, Stefan; Isaksen, Leif.....</i>	94
TEI P5 as a Text Encoding Standard for Multilevel Corpus Annotation <i>Bański, Piotr; Przepiórkowski, Adam.....</i>	98
Developing a Collaborative Online Environment for History – the Experience of British History Online <i>Blaney, Jonathan.....</i>	100
From Codework to Working Code: A Programmer's Approach to Digital Literacy <i>Bork, John.....</i>	101
Non-traditional Prosodic Features for Automated Phrase-Break Prediction <i>Brierley, Claire; Atwell, Eric.....</i>	103
How Do You Visualize a Million Links? <i>Brown, Susan; Antoniuk, Jeffery; Bauer, Michael; Berberich, Jennifer; Radzikowska, Milena; Ruecker, Stan; Yung, Terence.....</i>	105
Digital Libraries of Scholarly Editions <i>Buchanan, George; Bohata, Kirsti.....</i>	108

Digital Mediation of Modernist Literary Texts and their Documents	
<i>Byron, Mark</i>	110
Detection of Citations and Textual Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project	
<i>Büchler, Marco; Geßner, Annette; Heyer, Gerhard; Eckart, Thomas</i>	113
No Representation Without Taxonomies: Specifying Senses of Key Terms in Digital Humanities	
<i>Caton, Paul</i>	115
Modes of Seeing: Case Studies on the Use of Digitized Photographic Archives	
<i>Conway, Paul</i>	118
Digital Humanities Internships: Creating a Model iSchool-Digital Humanities Center Partnership	
<i>Conway, Paul; Fraistat, Neil; Galloway, Patricia; Kraus, Kari; Rehberger, Dean; Walter, Katherine</i>	120
Authorship Discontinuities of El Ingenioso Hidalgo don Quijote de la Mancha as detected by Mixture-of-Experts	
<i>Coufal, Christopher; Juola, Patrick</i>	123
Entropy and Divergence in a Modern Fiction Corpus	
<i>Craig, Hugh</i>	124
Objective Detection of Plautus' Rules by Computer Support	
<i>Deufert, Marcus; Blumenstein, Judith; Trebesius, Andreas; Beyer, Stefan; Büchler, Marco</i>	126
The ecology of longevity: the relevance of evolutionary theory for digital preservation	
<i>Doorn, Peter; Roorda, Dirk</i>	128
Joanna Baillie's : from Hypermedia Edition to Resonant Responses	
<i>Eberle-Sinatra, Michael; Crochunis, Tom C.; Sachs, Jon</i>	130
Does Size Matter? Authorship Attribution, Small Samples, Big Problem	
<i>Eder, Maciej</i>	132
Finding Stories in the Archive through Paragraph Alignment	
<i>Esteva, Maria; Xu, Weijia</i>	135
Naming the unnamed, speaking the unspoken, depicting the undepicted: story	
<i>Evans, Joanne; Morgan, Helen; Henningham, Nikki</i>	138
The Social Lives of Books: Mapping the Ideational Networks of Toni Morrison	
<i>Finn, Edward</i>	140
Codifica digitale e semiotica della cultura: un esperimento	
<i>Fiormonte, Domenico; Guadalupi, Laura</i>	143
Open vs. Closed: Changing the Culture of Peer Review	
<i>Fitzpatrick, Kathleen</i>	146
Using ODD for Multi-purpose TEI Documentation	
<i>Flanders, Julia; Bauman, Syd</i>	148
Xiakou: A Case Study in Digital Ethnography	
<i>Flower, John; Leonard, Pamela; Martin, Worthy</i>	150

Challenges of Linking Digital Heritage Scientific Data with Scholarly Research: From Navigation to Politics	
<i>France, Fenella G.; Toth, Michael B.; Hansen, Eric F.</i>	153
Building Dynamic Image Collections from Internet	
<i>Fu, Liuliu; Maly, Kurt; Wu, Harris; Zubair, Mohammad</i>	156
GIS, Texts and Images: New approaches to landscape appreciation in the Lake District	
<i>Gregory, Ian</i>	159
Capturing Visitor Experiences for Study and Preservation	
<i>Guy, Georgina; Dunn, Stuart; Gold, Nicolas</i>	160
The Diary of a Public Man: A Case Study in Traditional and Non-Traditional Authorship Attribution	
<i>Holmes, David I.; Crofts, Daniel W.</i>	163
Using the Universal Similarity Metric to Map Correspondences between Witnesses	
<i>Holmes, Martin</i>	165
Teasing Out Authorship and Style with T-tests and Zeta	
<i>Hoover, David L.</i>	168
A New Digital Method for a New Literary Problem: A Proposed Methodology for Bridging the "Generalist" - "Specialist" Divide in the Study of World Literature	
<i>Howell, Sonia; Keating, John G.; Kelleher, Margaret</i>	171
"Litmap": Networked Narratives	
<i>Hui, Barbara</i>	174
The Open Annotation Collaboration: A Data Model to Support Sharing and Interoperability of Scholarly Annotations	
<i>Hunter, Jane; Cole, Tim; Sanderson, Robert; Van de Sompel, Herbert</i>	175
A corpus approach to cultural keywords: a critical corpus-based analysis of ideology in the Blair years (1998-2007) through print news reporting	
<i>Jeffries, Lesley; Walker, Brian David</i>	178
The Modern Art Iraq Archive (MAIA): Web tools for Documenting, Sharing and Enriching Iraqi Artistic Expressions	
<i>Kansa, Sarah Whitcher; Shabout, Nada; Al-Bahloly, Saleem</i>	181
A Data Model for Digital Musicology and its Current State – The Music Encoding Initiative	
<i>Kepper, Johannes</i>	184
From Text to Image to Analysis: Visualization of Chinese Buddhist Canon	
<i>Lancaster, Lewis</i>	185
Crossing the Boundary: Exploring the Educational Potential of Social Networking Sites	
<i>Lang, Anouk</i>	187
Queste del Saint Graal: Textometry Platform on the Service of a Scholarly Edition	
<i>Lavrentiev, Alexei; Serge, Heiden; Yepdieu, Adrien</i>	190

The Graceful Degradation Survey: Managing Digital Humanities Projects Through Times of Transition and Decline <i>Nowviskie, Bethany; Porter, Dot</i>	192
LAP, LICHEN, and DASS – Experiences combining data and tools <i>Opas-Hänninen, Lisa Lena; Juuso, Ilkka; Kretzschmar, William A. Jr.; Seppänen, Tapio</i>	194
Re-linking a Dictionary Universe or the Meta-dictionary Ten Years Later <i>Ore, Christian-Emil; Ore, Espen S.</i>	196
Digital Resources for Art-Historical Research: Critical Approach <i>Rodríguez Ortega, Nuria</i>	199
Towards Hermeneutic Markup: An architectural outline <i>Piez, Wendell</i>	202
Works, Documents, Texts and Related Resources for Everyone <i>Robinson, Peter; Meschini, Federico</i>	206
A Day in the Life of Digital Humanities <i>Rockwell, Geoffrey; Ruecker, Stan; Organisciak, Peter; Meredith-Lobay, Megan; Kamal, Ranaweeram; Sinclair, Stéfan</i>	208
Letters, Ideas and Information Technology: Using digital corpora of letters to disclose the circulation of knowledge in the 17th century <i>Roorda, Dirk; Bos, Erik-Jan; van den Heuvel, Charles</i>	211
Pointless Babble or Enabled Backchannel: Conference Use of Twitter by Digital Humanists <i>Ross, Claire; Terras, Melissa; Warwick, Claire; Welsh, Anne</i>	214
The State of Non-Traditional Authorship Attribution Studies – 2010: Some Problems and Solutions <i>Rudman, Joseph</i>	217
Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? <i>Rybicki, Jan; Eder, Maciej</i>	219
Reading Darwin Between the Lines: A Computer-Assisted Analysis of the Concept of Evolution in <i>Sainte-Marie, Maxime B.; Meunier, Jean-Guy; Payette, Nicolas; Chartier, Jean-François</i>	225
The TEI's Extramural Journal Project: Exploring New Digital Environments and Defining a New Genre in Academic Publishing <i>Schlitz, Stephanie A.</i>	228
The Specimen Case and the Garden: Preserving Complex Digital Objects, Sustaining Digital Projects <i>Schlosser, Melanie; Ulman, H. Lewis</i>	230
A Tale of Two Cities: Implications of the Similarities and Differences in Collaborative Approaches within the Digital Libraries and Digital Humanities Communities <i>Siemens, Lynne; Cunningham, Richard; Duff, Wendy; Warwick, Claire</i>	232

Unfolding History with the Help of the GIS Technology: a Scholar-Librarian Quest for Creating Digital Collections <i>Smith, Natasha; Allen, Robert; Whisnant, Anne; Eckhardt, Kevin; Moore, Elise</i>	235
WW1 and WW2 on a Specialist E-forum. Applying Corpus Tools to the Study of Evaluative Language <i>Sokól, Małgorzata</i>	238
Visualization and Analysis of Visiting Styles in 3D Virtual Museums <i>Sookhanaphibarn, Kingkarn; Thawonmas, Ruck</i>	239
Two representations of the semantics of TEI Lite <i>Sperberg-McQueen, C. M.; Marcoux, Yves; Huitfeldt, Claus</i>	244
Thinking Archivally: Search and Metadata as Building Blocks for a New Digital Historiography <i>Sternfeld, Joshua</i>	246
e-Vocative Cases: Digitality and Direct Address <i>Swanstrom, Lisa</i>	249
Digitizing the Act of Papyrological Interpretation: Negotiating Spurious Exactitude and Genuine Uncertainty <i>Tarte, Ségolène M.</i>	251
Reimagining the Dictionary, or Why Lexicography Needs Digital Humanities <i>Tasovac, Toma</i>	254
Contexts, Narratives, and Interactive Visual Analysis of Names in the Japanese Hyohanki Diary <i>Toledo, Alejandro; Thawonmas, Ruck</i>	257
“Quivering Web of Living Thought”: Mapping the Conceptual Networks of Swinburne's <i>Walsh, John A.; Foong, Pin Sym; Anand, Kshitiz; Ramesh, Vignesh</i>	260
“It's Volatile”: Standards-Based Research & Research-Based Standards Development <i>Walsh, John A.; Hooper, Wally</i>	262
Quelques réflexions sur l'effet propédeutique des catalogues des collections des musées en ligne <i>Welger-Barboza, Corinne</i>	264
“Any more Bids?”: Automatic Processing and Segmentation of Auction Catalogs <i>West, Kris; Llewellyn, Clare; Burns, John</i>	267
Mandoku – An Incubator for Premodern Chinese Texts – or How to Get the Text We Want: An Inquiry into the Ideal Workflow <i>Wittern, Christian</i>	271
Towards a Computational Narration of Inner World <i>Zhu, Jichen</i>	273
Posters	
An Approach to Ancient-to-modern and Cross-script Information Access for Traditional Mongolian Historical Collections <i>Batjargal, Biligsaikhan; Khaltarkhuu, Garmaabazar; Kimura, Fuminori; Maeda, Akira</i>	279

A Digital Archive of Buddhist Temple Gazetteers <i>Bingenheimer, Marcus; Hung, Jen-jou</i>	282
Preparing the DARIAH e-Infrastructure <i>Blanke, Tobias; Haswell, Eric Andrew</i>	284
Cultures of Knowledge: An Intellectual Geography of the Seventeenth-Century Republic Letters <i>Brown, James; Hotson, Howard; Jefferies, Neil</i>	285
Supporting User Search for Discovering Collections of Interest <i>Buchanan, George; Dodd, Helen</i>	287
An Inter-Disciplinary Approach to Web Programming: A Collaboration Between the University Archives and the Department of Computer Science <i>Bunde, Janet Marie; Engel, Deena</i>	290
Citation Rhetoric Examined <i>Dobson, Teresa M.; Eberle-Sinatra, Michael; Ruecker, Stan; Lucky, Shannon</i>	292
Evidence of Intertextuality: Investigating Paul the Deacon's <i>Forstall, C. W.; Jacobson, S.L.; Scheirer, W. J.</i>	294
Historical Interpretation through Multiple Markup: The Case of Horatio Nelson Taft's Diary, 1861-62 <i>Garfinkel, Susan; Heckscher, Jurretta Jordan</i>	296
Diple, modular methodology and tools for heterogeneous TEI corpora <i>Glorieux, Frédéric; Canteaut, Olivier; Jolivet, Vincent</i>	299
A New Spatial Analysis of the Early Chesapeake Architecture <i>Graham, Wayne</i>	301
The Importance of Pedagogy: Towards a Companion to Teaching Digital Humanities <i>Hirsch, Brett D.; Timney, Meagan</i>	303
A Bilingual Digital Edition of Trinity College Cambridge MS O.1.77. <i>Honkapohja, Alpo</i>	304
The Craig Zeta Spreadsheet <i>Hoover, David L.</i>	306
The Dickens Lexicon and its Practical Use for Linguistic Research <i>Hori, Masahiro; Imabayashi, Osamu; Tabata, Tomoji; Nishio, Miyuki</i>	309
Dingler-Online – The Digitized "Polytechnisches Journal" on Goobi Digitization Suite <i>Hug, Marius; Kassung, Christian; Meyer, Sebastian</i>	311
The MLCD Overlap Corpus (MOC) <i>Huitfeldt, Claus; Sperberg-McQueen, C. M.; Marcoux, Yves</i>	313
Creative Engagement with Creative Works: a New Paradigm for Collaboration <i>Jones, Steven E.; Shillingsburg, Peter; Thiruvathukal, George K.</i>	317
Distant Reading and Mapping Genre Space via Conjecture-based Distance Measures <i>Juola, Patrick</i>	319

Psycholinguistically Plausible Events and Authorship Attribution <i>Juola, Patrick</i>	320
National Digital Library of Finland: Putting the Resources of Culture, Science and Teaching at Everyone's Fingertips <i>Kautonen, Heli; Sainio, Tapani; Vakkari, Mikael</i>	321
Towards Digital Built Environment Studies: An Interface Design for the Study of Medieval Delhi <i>Keshani, Hussein</i>	324
Prop Revisited: Integration of Linguistic Markup into Structured Content Descriptors of Tales <i>Lendvai, Piroska; Declerck, Thierry; Darányi, Sándor; Malec, Scott</i>	327
Extracting domain knowledge from tables of contents <i>Lüngen, Harald; Lobin, Henning</i>	331
Museums of the virtual future <i>Martinet, Marie-Madeleine; Gallet-Blanchard, Liliane</i>	335
Discursive Metadata and Controlled Vocabularies <i>Mylonas, Elli; Wendts, Heidi; Bodel, John</i>	337
The Digital Ark: From Taxonomy to Ontology in 17th-century Collections of Curiosities <i>Nelson, Brent</i>	339
"Inventing the Map:" from 19th-century Pedagogical Practice to 21st-century Geospatial Scholarship <i>Nowviskie, Bethany</i>	341
An Open Source Toolkit for Flexible Browsing of Historical Maps on the Web <i>Ohno, Shin; Saito, Shinya; Inaba, Mitsuyuki</i>	344
Text-Image linking of Japanese historical documents: Sharing and exchanging data by using text-embedded image file <i>Okamoto, Takaaki</i>	347
Knowledge and Conservation - Creating the Digital Library of New Hispanic Thought <i>Priani, Ernesto; Galina, Isabel; Martínez, Alí; Chávez, Guillermo</i>	350
Digital Forensics, Textual Criticism, and the Born Digital Musical <i>Reside, Doug</i>	353
Literary Theory and Theatre Practice: A Comparative Study of Watching the Script and the Simulated Environment for Theatre <i>Roberts-Smith, Jennifer; Dobson, Teresa M.; Gabriele, Sandra; Ruecker, Stan; Sinclair, Stéfan; Bouchard, Matt; DeSouza-Coelho, Shawn; Kong, Annemarie; Lam, David; Rodriguez, Omar; Taylor, Karen</i>	354
The Person Data Repository <i>Roeder, Torsten</i>	356
Structured and Unstructured: Extracting Information from Classics Scholarly Texts <i>Romanello, Matteo</i>	360

Original, Translation, Inflation. Are All Translations Longer than Their Originals? <i>Rybicki, Jan</i>	363
A Platform for Cultural Information Visualization Using Schematic Expressions of Cube <i>Saito, Shinya; Ohno, Shin; Inaba, Mitsuyuki</i>	365
Generation of Emotional Dance Motion for Virtual Dance Collaboration System <i>Seiya, Tsuruta; Woong, Choi; Kozaburo, Hachimura</i>	368
“You don't have to be famous for your life to be history”: The Dusenberry Journal and img2xml <i>Smith, Natasha; Cayless, Hugh</i>	372
Delivering virtual reality: a proposal for facilitating pedagogical use of three-dimensional computer models of historic urban environments <i>Snyder, Lisa M.; Friedman, Scott</i>	373
Digitizing Ephemera and Parsing an 1862 European Itinerary <i>Tomasek, Kathryn; Stickney, Zephorene L</i>	377
Critical Editing of Music in the Digital Medium: an Experiment in MEI <i>Viglianti, Raffaele</i>	380
LogiLogi: The Quest for Critical Mass <i>Wiersma, Wybo</i>	383
Software Demonstration, “Emergent Time” timeline tool <i>York, Christopher; Trettien, Whitney</i>	386
Putting Edmonton on the (Google) Map <i>Zwicker, Heather; Engel, Maureen</i>	387
Text Encoding and Ontology – Enlarging an Ontology by Semi-Automatic Generated Instances <i>Zöllner-Weber, Amélie</i>	390

List of reviewers

- Akama, Dr. Hiroyuki
- Anderson, Dr. Deborah
- Anderson, Jean Gilmour
- Andreev, Vadim Sergeevich
- Baayen, Prof. Rolf Harald
- Barney, Brett
- Bauman, Syd
- Baumann, Ryan Frederick
- Bearman, David
- Beavan, David
- Bellamy, Dr. Craig
- Bennis, Prof. Hans
- Bentkowska-Kafel, Dr. Anna
- Bia, Prof. Alejandro
- Biber, Dr. Hanno
- Blanke, Dr. Tobias
- Bodard, Dr. Gabriel
- Bodenhamer, Dr. David
- Bol, Prof. Peter Kees
- Booij, Prof. Geert E.
- Boot, Peter
- Bosse, Arno
- Boves, Prof. Lou
- Bowen, Prof. William
- Bradley, John
- Brey, Gerhard
- Brown, Prof. Susan
- Burnard, Lou
- Burr, Prof. Elisabeth
- Bush, Chuck
- Caton, Dr. Paul
- Cayless, Dr. Hugh
- Chen, Szu-Pei
- Chesley, Paula Horwath
- Ciula, Dr. Arianna
- Clement, Dr. Tanya
- Conner, Prof. Patrick
- Connors, Louisa
- Cooney, Dr. Charles M.
- Cooper, Dr. David Christopher
- Cossard, Prof. Patricia Kosco
- Craig, Prof. Hugh
- Cummings, Dr. James C.
- Cunningham, Dr. Richard
- Dahlstrom, Dr. Mats
- David, Stefano
- Dawson, Dr. John
- Devlin, Dr. Kate
- DiNunzio, Joseph
- Dombrowski, Quinn Anya
- Downie, Prof. J. Stephen
- Dunn, Dr. Stuart
- Durand, Dr. David G.
- Durusau, Patrick
- Edmond, Dr. Jennifer C
- Egan, Dr. Gabriel
- Eide, Øyvind
- Ell, Dr. Paul S
- Esteva, Dr. Maria
- Everaert, Prof. Martin
- Fiormonte, Dr. Domenico
- Fischer, Dr. Franz
- Fitzpatrick, Prof. Kathleen
- Flanders, Dr. Julia
- Flatscher, Markus
- Forest, Dr. Dominic
- Fraistat, Prof. Neil R.
- France, Dr. Fenella Grace
- French, Dr. Amanda
- Fritze, Christiane
- Furuta, Dr. Richard
- Galina Russell, Dr. Isabel
- Gallet-Blanchard, Prof. Liliane
- Gants, Prof. David
- Gärtner, Prof. Kurt
- Gartner, Richard
- Gilbert, Joseph
- Giordano, Dr. Richard
- Gow, Ann
- Groß, Dr. Nathalie

- Gueguen, Gretchen Mary
- Hanlon, Ann
- Hanrahan, Dr. Michael
- Hawkins, Kevin Scott
- Hernández Figueroa, Prof. Zenón
- Hirsch, Dr. Brett
- Hockey, Prof. Susan
- Holmes, Martin
- Hoover, Prof. David L.
- Horton, Prof. Tom
- Hswe, Patricia
- Hughes, Lorna
- Huitfeldt, Claus
- Hulk, Prof. Aafke
- Hunyadi, Prof. László
- Isaksen, Leif
- Ivanovs, Prof. Aleksandrs
- Jessop, Martyn
- Jockers, Dr. Matthew
- Johnsen, Dr. Lars
- Johnson, Dr. Ian R.
- Juola, Prof. Patrick
- Kaislaniemi, Samuli
- Kansa, Dr. Sarah Whitcher
- Khosmood, Foaad
- Kirschenbaum, Prof. Matthew
- Knowles, Prof. Anne Kelly
- Kraus, Dr. Kari Michaele
- Krauwer, Steven
- Kretzschmar, Dr. William
- Krot, Michael Adam
- Lancaster, Dr. Lewis Rosser
- Lavagnino, Dr. John
- Lavrentiev, Dr. Alexei
- Leitch, Caroline
- Lewis, Benjamin G.
- Llewellyn, Clare
- Luyckx, Kim
- Mahony, Simon
- Makinen, Dr. Martti
- Martin, Prof. Worthy N.
- Martinet, Prof. Marie-Madeleine
- McCarty, Prof. Willard
- McDaniel, Dr. Rudy
- Meister, Prof. Jan Christoph
- MendezRodriquez, Dr. Eva
- Meschini, Federico
- Miyake, Dr. Maki
- Mostern, Dr. Ruth
- Mylonas, Elli
- Myojo, Dr. Kiyoko
- Nagasaki, Kiyonori
- Nelson, Prof. Brent
- Nerbonne, Prof. John
- Neuman, Dr. Michael
- Newton, Greg T.
- Nowviskie, Dr. Bethany
- O'Donnell, Dr. Daniel Paul
- Olsen, Prof. Mark
- Opas-Hänninen, Prof. Lisa Lena
- Ore, Dr. Christian-Emil
- Ore, Espen S.
- Parker, Alexander
- Pasanek, Brad
- Pierazzo, Dr. Elena
- Piez, Dr. Wendell
- Pitti, Daniel
- Pytlak Zillig, Prof. Brian L.
- Rahtz, Sebastian
- Rains, Michael John
- Ramsay, Dr. Stephen
- Rehbein, Dr. Malte
- Rehm, Dr. Georg
- Renear, Dr. Allen H.
- Reside, Dr. Doug
- Robertson, Prof. Bruce
- Robey, Prof. David
- Robinson, Prof. Peter
- Rockwell, Prof. Geoffrey
- Rodríguez, Dr. Nuria
- Roe, Glenn H.
- Romary, Prof. Laurent

- Roueché, Prof. Charlotte
- Roued-Cunliffe, Henriette
- Rudman, Prof. Joseph
- Ruecker, Dr. Stan
- Russo, Dr. Angelina
- Rybicki, Dr. Jan
- Saint-Dizier, Prof. Patrick
- Sanz, Prof. Concha
- Schlitz, Dr. Stephanie
- Schmidt, Harry
- Schmidt, Sara A.
- Schreibman, Prof. Susan
- Sculley, D.
- Seppänen, Prof. Tapio
- Shaw, William Stewart
- Shawver, Dr. Gary
- Siemens, Dr. Lynne
- Siemens, Dr. Raymond George
- Simons, Prof. Gary F.
- Sinclair, Prof. Stéfan
- Singer, Kate
- Smith, Natalia (Natasha)
- Snyder, Dr. Lisa M.
- Spence, Paul
- Sperberg-McQueen, Dr. Michael
- Spiro, Dr. Lisa
- Steggie, Prof. Matthew
- Sternfeld, Dr. Joshua
- Stokes, Dr. Peter A.
- Sukovic, Suzana
- Suzuki, Dr. Takafumi
- Swanstrom, Dr. Elizabeth Anne
- Tabata, Prof. Tomoji
- Terras, Dr. Melissa
- Thaller, Prof. Manfred
- Tripp, Mary L.
- Tufis, Prof. Dan
- Unsworth, John
- Van den Branden, Ron
- Van den Herik, Prof. H. J.
- Van Elsacker, Bert
- Vanhoutte, Edward
- Váradi, Dr. Tamás
- Walker, Brian David
- Walsh, Prof. John
- Warwick, Dr. Claire
- Webb, Sharon
- Wiesner, Dr. Susan L.
- Wilkens, Dr. Matthew
- Willett, Perry
- Winder, Dr. William
- Witt, Dr. Andreas
- Wittern, Prof. Christian
- Wolff, Prof. Mark
- Worthey, Glen
- Yu, Dr. Bei
- Zafrin, Dr. Vika
- Zhang, Junte
- Zimmerman, Matthew

Plenary Sessions

To Hold Up a Mirror: Preservation and Interpretation of Performance in a Digital Age

Henry, Charles J.

Council on Library and Information Resources (CLIR)

It is commonplace to separate the methods of preservation of our cultural heritage from scholarly interpretation. Preservation is often described in more technical terms, while scholarship is deemed an intellectual engagement removed from the the technicalities of electronic capture and persistence. This presentation challenges that distinction, and rather explores the dynamic, causal relationship between preserving a performance event and its subsequent interpretation.

The scholar's reception and elucidation of performance can be traced back at least to Aristarchus and his collation and annotation of the various written records of recitations of Homer's epic poetry that had accumulated by the second century B.C.E. More recently, the digitization of the Bayeux Tapestry illuminates the interplay between the translation, from one medium to another, and preservation of a fundamentally important object of human expression and its determining influence on how that object may be interpreted and received subsequent to its digitization.

Today, performance often entails rich, multimedia elements that pose considerable difficulties for preserving the event and making it accessible over time. As importantly, the methods of capture can limit but also allow new and exciting opportunities for scholarly exegesis, including the capture of various stages and components of the creative process, illuminating the context and history of the performance 'event'.

What does it mean to preserve our cultural record digitally? What new methods of interpretation may arise in response to a digital record of an otherwise fleeting and ephemeral event? What new means of publication will be needed to communicate adequately the various 'readings' of a digitally preserved performance?

Humanities Computing in an Age of Social Change

Raben, Joe

Queens College of the City University of New York

Humanities computing in the United States can be considered to have started a few years before 1964, when IBM sponsored what it designated as a Literary Data Processing Conference. While most of the participants in that early conference, despite their often-expressed interest in the concept of nonlinear visualization of texts, were clearly oriented toward the goal of producing a printed book, two generations after that founding conference, we can recognize that the value of their work lies in their having begun to establish humanities computing as a valid occupation of scholars. There was, nevertheless, a need for a common ground on which to record and exchange our ideas of where this new mode of scholarship was leading us; hence the print journal *Computers and the Humanities*.

Not evident to that handful of pioneers in 1964 was the amazing growth of computer applications throughout society that were made possible by the technological advances of the next half-century. Humanities computing has advanced as far as it has almost exclusively because of the revolution on the technological side. Because of the computer revolution, the world we inhabit is no more like the one known to previous generations than that of the twentieth century resembled any of its predecessors.

The drastic changes in our world in the almost half-century since 1964 makes clear that we can no more predict the changes to come than those pioneers did in their own time. Computer-based communication, in particular and especially in its printed form, is being violently altered by the new technology. The openness of the Internet and the Web being a manifestation of a democratic spirit, the burgeoning role of computers in education, including humanities education, can only continue to disrupt the traditional structure of academe.

The long-term consequences of the increasing cost of a postsecondary education and the increasing availability of resources that exceed those of any university would seem to drive toward the replacement of the bricks-and-mortar university by a totally online facility. That paradigm shift requires the aggregation in a central online location of information about the growing resources of the digital humanities. The generation that will prevail in the middle of the twenty-first century, wired to computers for all their needs, social as well as intellectual, will look beyond our current concepts of humanities. How well we prepare for that world, what foundation we construct to emphasize the positive potentials of whatever technology will have evolved, will be the measure of how much we have learned from our humanistic concern for our own history.

Present, Not Voting: Digital Humanities in the Panopticon

Terras, Melissa

Senior Lecturer in Electronic Communication, Deputy Director of UCL Centre for Digital Humanities, Department of Information Studies

The field of Digital Humanities continues to change and evolve rapidly, as we utilise, appropriate, and develop internet, communication, and computational technologies. In this plenary, a specific focus will be placed on one individual project – The Transcribe Bentham project at UCL (<http://www.ucl.ac.uk/transcribe-bentham/>) – as a viewpoint through which to witness the changing demands and needs placed on those working within the Digital Humanities.

Transcribe Bentham is a one year, Arts and Humanities Research Council funded project, housed under the auspices of the Bentham Project at UCL (<http://www.ucl.ac.uk/Bentham-Project/>). The Bentham Project aims to produce new editions of the scholarship of Jeremy Bentham (1748-1832). Bentham was a political radical. An English jurist, philosopher, and legal and social reformer, he became a leading theorist in Anglo-American philosophy of law, and influenced the development of welfarism. He is well known for his advocacy of utilitarianism and animal rights, but is perhaps most famous for his work on the “panopticon”: a type of prison in which wardens can observe (-*opticon*) all (*pan-*) prisoners without the incarcerated being able to tell whether or not they are being watched. This concept has influenced prison reform, philosophy, literature, and social media since.

Bentham and UCL have a close relationship. Whilst it is untrue that he founded UCL, he did influence those who did, and as the first English University to open its doors to all, regardless of race, creed or political belief (provided they could afford the fees) UCL went a long way to fulfilling Bentham's vision of how Universities should operate. He took a great interest in the new institution, and UCL now hosts Bentham's 60,000 pages of handwritten manuscripts arranged in 174 boxes. His “auto-icon” famously sits in the main cloisters: Bentham's preserved skeleton, dressed in his own clothes, surmounted by a wax head, as per his will and testament. An apocryphal story has it that Bentham's “auto-icon” is wheeled yearly into UCL Senate meetings, where he is noted in the minutes as being “present, not voting”.

Twelve volumes of Bentham's correspondence have so far been published by the Bentham Project, plus various collections of his work on jurisprudence and legal matters (<http://www.ucl.ac.uk/Bentham-Project/Publications/index.htm>). However, there is much more work to be done to make his writings more accessible, and to provide transcripts of the materials therein. Although a previous grant from the AHRC in 2003-6 has allowed for the completion of a catalogue of the manuscripts held within UCL (<http://www.benthampapers.ucl.ac.uk/>), and transcriptions have been completed of some 10,000 folios, there are many hours of work that need to be invested in reading, transcribing, labelling, and making accessible the works of this interdisciplinary historical figure if they are to be analysed, consulted, and utilised by scholars across the various disciplines interested in Bentham's writings.

Crowdsourcing – the harnessing of online activity to aid in large scale projects that require human cognition – is becoming of interest to those in the library, museum and cultural heritage industry, as institutions seek ways to publically engage their online communities, as well as aid in creating useful and usable digital resources. As one of the first cultural and heritage projects to apply crowdsourcing to a non-trivial task, UCL's Bentham Project has recently set up the "Transcribe Bentham" initiative; an ambitious, open source, participatory online environment which is being developed to aid in transcribing 10,000 folios of Bentham's handwritten documents. To be launched in July 2010, this experimental project will aim to engage with individuals such as school children, amateur historians, and other interested parties, who can provide time to help us read Bentham's manuscripts. The integration of user communities will be key to the success of the project, and an additional project remit is to monitor the success of trying to engage the wider community with such documentary material: will we get high quality, trustworthy transcriptions as a result of this work? Will people be interested in volunteering their time and effort to

read the (poor) handwriting of a great philosopher? What technical and pragmatic difficulties will we run into? How can we monitor success in a crowdsourced environment?

In addition to introducing the project, this plenary will use the Bentham Project, and the Transcribe Bentham open participatory initiative, as a panopticon through which to view the changing nature of Digital Humanities scholarship, and the role of the Digital Humanities scholar, alluding to other related work in the field, and concerns which have emerged in the development and discussion of the project aims. How do issues such as digital identity, scholarly and community engagement, professionalization and employment issues, funding, short scale projects, and large scale digitisation affect the role of the digital humanist? How can we successfully embrace ever changing internet technologies, and trends such as crowdsourcing, to further our research in cultural heritage and computational methods? How can we measure the impact of our endeavours in the online environment, and how can we persuade others of the value of our work? What can we do better – as a discipline – to make sure our voice is heard in the academy, so that the role, remit, and focus of Digital Humanities can be treated as a bona fide academic endeavour? What can we do better – as individuals – to further the cause of Digital Humanities as a discipline, as well as fostering our own scholarly development?

By looking at specific issues that have arisen in the Transcribe Bentham project, this plenary aims to provide an honest and open overview of pressing concerns and opportunities for both the project, and Digital Humanities as a discipline, in the current information and academic environments.

Melissa Terras and Transcribe Bentham can be followed on twitter at #TranscriBentham and #melissaterras respectively. The complete transcript of this plenary will be made available shortly after delivery on Melissa Terras' blog: <http://melissaterras.blogspot.com/>, and viewable thereafter in the blog archive for July 2010.

Art Installations

Vanishing Point(s) and Communion

Michael Takeo Magruder

King's College London

Denard, Hugh

King's College London

1. Vanishing Point(s)

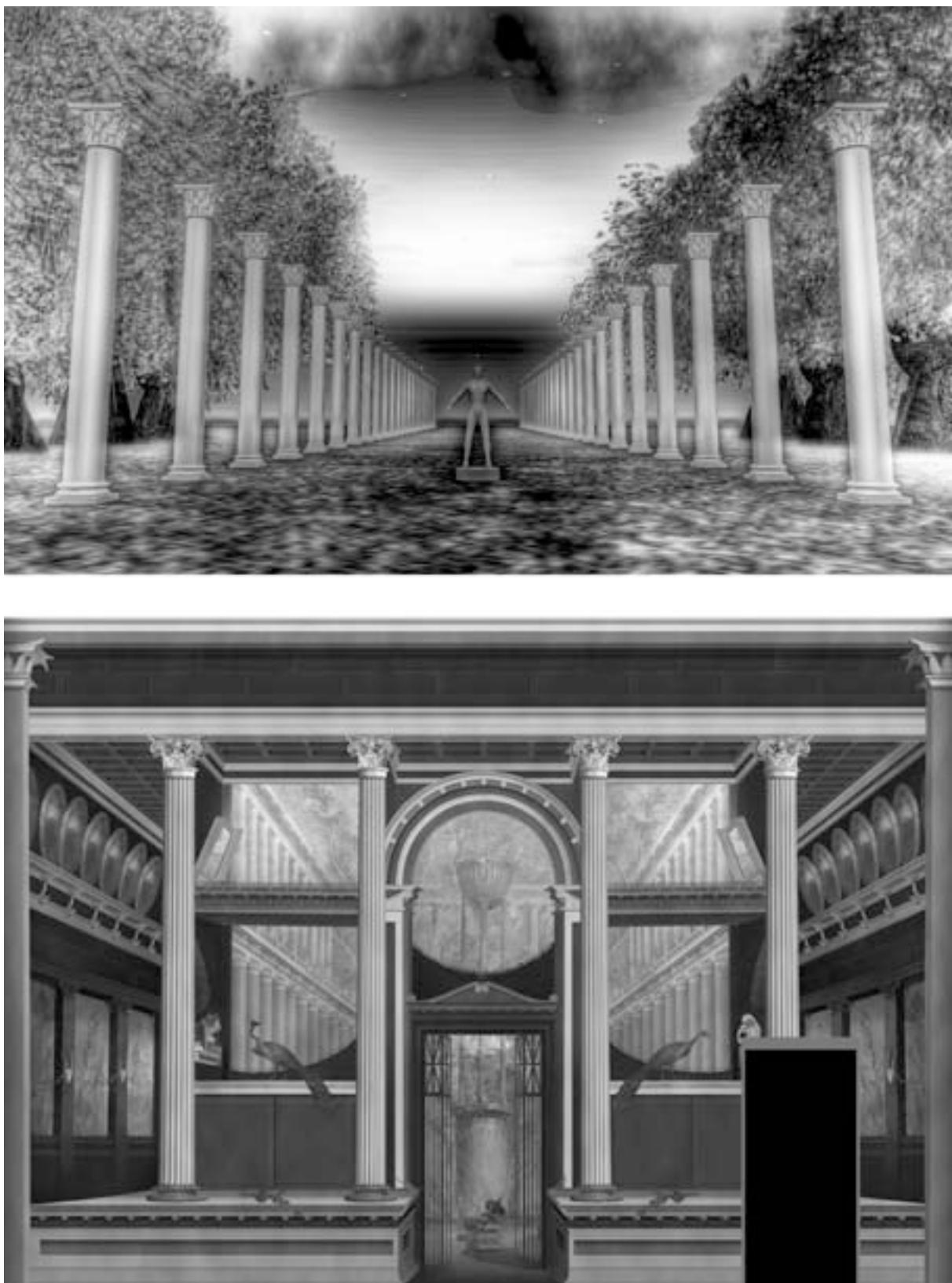
Vanishing Point(s) is a site-specific art installation by Michael Takeo Magruder and Dr. Hugh Denard that explores creative collisions and collaborative possibilities between contemporary art discourse and humanities research. Commissioned for Digital Humanities 2010 for the Great Hall of the Grade I listed King's Building created in 1831 by English architect Sir Robert Smirke (1781 - 1867), the project conjoins Takeo's long-standing use of computational processes and virtual environments as frameworks for artistic expression and Denard's studies of the playfully illusionistic and fantastical worlds of Roman fresco art.

Vanishing Point(s) takes as its inspiration the astonishingly complex and beautiful ways in which Roman architecture and painting often converged, immersing the viewer in imagined spaces – idealised cities and gardens, palaces and shrines, theatres and basilicas – and beguilingly interweaving physical architecture with painted views so that it is not always easy to discern fact from fantasy; these are indeed "virtual" worlds that speak to the digitally-generated virtual worlds of the Avatar Age. The creators have drawn deeply upon the conceptual and compositional principles of theatrically-inspired Roman frescoes to form structured vistas in the online synthetic realm of Second Life as the visual source material for a new work that evokes the spatial-pictorial tradition of medieval stained glass.

Through processes of reimagining and reconstructing, **Vanishing Point(s)** invites ancient and medieval principles of spatial and pictorial representation to speak to this present age of data networks, mixed-realities and multimodal existence.

Michael Takeo Magruder (b.1974, US/UK) BSc (University of Virginia) is an artist and researcher based in King's Visualisation Lab, King's College London. His creative practice extends traditional modes of artistic production through the use of emerging media including high-performance computing, mobile devices and virtual environments, blending Information Age technologies and aesthetics to explore the networked, digital world. His work has been showcased in over 200 exhibitions in 30 countries, including Manifesta 8: European Biennial of Contemporary Art, Murcia; the Courtauld Institute of Art, London; EAST International 2005, Norwich; Georges Pompidou Center, Paris; Tokyo Metropolitan Museum of Photography, Japan; and Trans-Media-Akademie, Hellerau. His artistic practice has been funded by the Esmée Fairbairn Foundation, the Andy Warhol Foundation for the Visual Arts, Arts Council England and The National Endowment for the Arts, USA, and he has been commissioned by numerous public galleries in the UK and abroad and by the leading Internet Art portal <http://turbulence.org>.

Dr. Hugh Denard (b.1970, IE) BA (Dublin), MA, PhD (Exeter) is Associate Director of King's Visualisation Lab in the Centre for Computing in the Humanities, King's College London, where he convenes the MA in Digital Culture and Technology. He is a specialist in ancient Greek, Roman and twentieth-century Irish theatre history and in the application of digital visualisation methodologies in the humanities. He proposed and edits the internationally-recognised *London Charter for the Computer-based Visualisation of Cultural Heritage* and has jointly directed numerous projects funded by the AHRC, Arts Council England, British Council-Italian Ministry of Research, Eduserv Foundation, JISC, Leverhulme Trust and the Metropolitan Museum of Art. In Second Life, he co-directed the Theatron 3 project, collaborates with artist Michael Takeo Magruder, curates Digital Humanities Island, and teaches postgraduate modules on applied visualisation in the arts, humanities and cultural heritage.



(top) Work-in-progress study for ***Vanishing Point(s)***, 2010.

(bottom) Visual reconstruction of the Villa of Olpontis, *oecus* Room 15, east wall fresco by Martin Blazeby, KVL, 2010.

2. Communion

Communion is new media art installation by Michael Takeo Magruder that reflects upon aesthetic and informational qualities of language within today's technologically-enabled and multicultural society. The

artwork is created exclusively from sampled 'front-pages' of the BBC's online international news service, digitally recorded at a finite moment in time. Each of the composition's forty distinct elements correlates to a different language edition of the BBC website. The captured web-pages have been algorithmically processed through a single, predefined sequence of instructions. The resulting images are visually reminiscent of traditional stained-glass windows and Rorschach inkblots, and through such spiritual and psychological references, notions of the transcendental and mechanisms for emotional response are introduced. The visual structures are only semi-abstract and even though the media itself provides the aesthetic essence of the work, the language – and information it contains – is still partially discernible.

Originally commissioned in 2005 by Arts Council England for the forty leaded-glass windows in the main gallery space of 20-21 Visual Arts Centre, North Lincolnshire, **Communion** will be reworked as a series of wall-mounted prints for Digital Humanities 2010.



Installation view of **Communion** at 20-21 Visual Arts Centre, UK, 2005.



Detailed view of ***Communion*** (Spanish), 2005.

The Embroidered Digital Commons: Rescension

Carpenter, Ele

Goldsmiths College, University of London

The 'Embroidered Digital Commons' is an artwork facilitated by Ele Carpenter as part of the Open Source Embroidery project, utilising social and digital connectivity. The artwork is a practice-based research project exploring the language of the digital commons through close reading and stitching, in which conference delegates are invited to participate.

In 2003 the Raqs Media Collective wrote *A Concise Lexicon of/for the Digital Commons*. The full lexicon is an A-Z of the interrelationship between social, digital and material space. It weaves together an evolving language of the commons that is both poetic and informative. The terms of the lexicon are: Access, Bandwidth, Code, Data, Ensemble, Fractal, Gift, Heterogeneous, Iteration, Kernel, Liminal, Meme, Nodes, Orbit, Portability, Quotidian, Rescension, Site, Tools, Ubiquity, Vector, Web, Xenophilly, Yarn, and Zone.

The 'Embroidered Digital Commons' is an ambitious project to hand-embroider the whole lexicon, term by term, through workshops and events as a practical way of close-reading and discussing the text and its current meaning. Each term is chosen in relation to the specific context of its production through group workshops, conferences and events. The term 'Yarn' was embroidered at the HUMlab Digital Humanities Media Lab at Umeå University in Sweden, 2009. Here at the DH2010 conference we will aim to stitch the complex term 'Rescension'.

The concept of the digital commons is based on the potential for everything that is digital to be common to all. Like common grazing land, this can mean commonly owned, commonly accessed or commonly available. But all of these blurred positions of status and ownership have complex repercussions in the field of intellectual property and copyright. The commons has become synonymous with digital media through the discourse surrounding free and open source software and creative commons licensing. The digital commons is a response to the inherent 'copy n paste' reproducibility of digital data, and the cultural forms that they support. Instead of trying to restrict access, the digital commons invite open participation in the production of ideas and culture - where culture is not something you buy, but something you do.

The use of metaphor to explain technological concepts was expertly developed by Lady Ada Byron Lovelace in her letters and notes accompanying the Analytical Engine. Her love of poetical science combined the influences of her father, Lord Byron, and her mathematical mother, Lady Lovelace. Ada gave us the textile - metaphors for code and programming in the 1830s, informed by the binary punch card programming of the Jacquard Loom and Charles Babbage's Analytical Engine (Plant, 1997).

Rescension

The project for the DH2010 conference is to consider and embroider the following text: "Rescension A re-telling, a word taken to signify the simultaneous existence of different versions of a narrative within oral, and from now onwards, digital cultures. Thus one can speak of a 'southern' or a 'northern' rescension of a myth, or of a 'female' or 'male' rescension of a story, or the possibility (to begin with) of Delhi/Berlin/Tehran 'rescensions' of a digital work. The concept of rescension is contraindicative of the notion of hierarchy. A rescension cannot be an improvement, nor can it connote a diminishing of value. A rescension is that version which does not act as a replacement for any other configuration of its constitutive materials. The existence of multiple rescensions is a guarantor of an idea or a work's ubiquity. This ensures that the constellation of narrative, signs and images that a work embodies is present, and waiting for iteration at more than one site at any given time. Rescensions are portable and are carried within orbiting kernels within a space. Rescensions taken together constitute ensembles that may form an interconnected web of ideas, images and signs." (Raqs Media Collective, 2003)

The embroidery is a rescension of the lexicon. As we sew we retell the story of the digital commons (itself creative commons licensed to be retold). And as we emphasise the line of our stitches and falter

over knotted words, we make our own subtle interpretations of the text, adding nuances of colour, and personalized references.

The Digital Humanities is a large net woven by many scholars from many fields, each with their own perspectives on the concept of how the digital is common to all people, and how it is restricted. Curiously we watch people sewing in the dim light – what does it say? What does it mean? Where are you from? How do you retell the story of your knowledge? What is your lexicon?

A Concise Lexicon of/for the Digital Commons uses its own terms to describe new terms. So the word ‘rescension’ is described in relation to ‘fractal’, ‘kernel’, ‘node’, and ‘ubiquity’.

In the ‘Embroidered Digital Commons’, the text forms a kernel which is “the central rescension, of a narrative, a code, a set of signs ... that invites modification, extrapolation and interpretation, by its very presence.” (Raqs Media Collective, 2003). It is the core of an idea at the centre of discourse.

The text is open for any group to embroider, as a fractal or fragment of “free cultural code” as described in the lexicon: a fractal “is a rescension of every other fractal that has grown from within it. In the same way a fragment of free code, or free cultural code, carries within it myriad possibilities of its own reproduction and dispersal within a shared symbolic or information space.” (Raqs Media Collective, 2003)

We are nodes in the network of communication, and each thread and word is reinforced through repetition, where every utterance is both the same and different each time. The lexicon describes “echoes and resonances” as “rescensions” which travel through nodes, “and each node is ultimately a direct rescension of at least one other node in the system and an indirect rescension of each junction within a whole cluster of other nodes.” (Raqs Media Collective, 2003). The embroidered patches become a central node, where ideas and arguments can cluster.

According to the Lexicon ‘Ubiquity’ is: “A rescension, when in orbit, crosses the paths of its variants. The zone where two orbits intersect is usually the site of an active transaction and transfer of meanings. Each rescension carries into its own trajectory memes from its companion. In this way, through the encounters between rescensions, ideas spread, travel and tend towards ubiquity.” (Raqs Media Collective, 2003)

By the end of the conference the term ‘Rescension’ will become ubiquitous: in our minds, in our presentations, in our conversations, in our pricked fingers and in our notes scribbled incomprehensibly in the dark. And possibly in a patchwork which defines or redefines the term, and enables it to continue traveling.

References

Plant, Sadie (1997). *Zeros and Ones: Digital Women and the New Technoculture*. London: Fourth Estate.

Raqs Media Collective (2003). 'A Concise Lexicon of/for the Digital Commons'. *Sarai Reader 03: Shaping Technologies*. Monica Narula, Shuddhabrata Sengupta, Jeebesh Bagchi, Ravi Vasudevan, Ravi Sundaram, Geert Lovink (eds.). Delhi/Amsterdam: Sarai-CSDS/WAAG, pp. 365. <http://www.raqsmedia.org/texts4.html>.

Ele Carpenter is a curator, artist and writer based in the UK and Sweden. She is a lecturer in MFA Curating at Goldsmiths College University of London, and is a Postdoctoral Research Fellow at HUMlab in affiliation with Bildmuseet at the University of Umeå, Sweden.

Since 2005 Ele has facilitated the Open Source Embroidery project using embroidery and code as a tool to investigate participatory production and distribution methods. Ele is currently working on the 'Open Source Crafter' publication, and facilitating the 'Embroidered Digital Commons' distributed embroidery.

Ele received her PhD on the relationship between politicised socially engaged art and new media art, with CRUMB at the University of Sunderland in 2008; and was previously Curator, NGCA Sunderland (1997-2002); Associate Curator, CCA Glasgow (2003-5).

Pre-conference Workshops

Access to the Grid: Interfacing the Humanities with Grid Technologies

Dunn, Stuart

stuart.dunn@kcl.ac.uk

King's College London

There can be little doubt that large-scale Grid infrastructures have transformed the way research is done in some parts of the physical sciences. High-profile enterprises such as the Large Hadron Collider would be of little use without computational infrastructures which are capable of supporting the vast quantities of data they produce. Although these branches of science are unique in terms of the volumes of data they contend with, other fields are encountering equivalent research problems which require Grid services and resources, and cognate technologies, tailored according to their own disciplinary needs. The humanities are no exception: recent engagement between the humanities and 'e-Science' (e.g. <http://www.ahessc.ac.uk/initiative-projects>) have shown that their complex data and research processes can be supported and enhanced using Grids and associated technologies. This workshop will seek to scope practical points of engagement, both current and potential, between Grid infrastructures in the UK and Europe. It will place particular emphasis on the portals and interface technology that humanists need in order to use Grids. The event will bring together leading European practitioners of digital humanities (many of whom have already done significant work with Grid infrastructures) together with representatives of key Grid infrastructure organizations, including EGI and the NGS. It will attempt to produce a roadmap of which areas of the humanities have most to gain from using Grids, and which do not; and how the Grid and humanities research communities can better work together in the future.

This event comes within the context of a major change in European research e-infrastructure. The Enabling Grids for E-Science (EGEE) will be replaced by the European Grid Initiative (<http://web.eu-eu.org/>). Given the latter's emphasis on federating National Grid Initiatives (NGIs), it is important that the digital humanities position themselves to gain maximum advantage both nationally and at a European level.

This workshop is being held with the support of JISC
(<http://www.jisc.ac.uk>)

Full day workshop: Monday, 5 July

The workshop programme is available online: <http://ahessc.ac.uk/grid-workshop-programme>

Content, Compliance, Collaboration and Complexity: Creating and Sustaining Information

Evans, Joanne

joanne.evans@unimelb.edu.au
University of Melbourne, Australia

Henningham, Nikki

n.henningham@unimelb.edu.au
University of Melbourne, Australia

Morgan, Helen

helen.morgan@unimelb.edu.au
University of Melbourne, Australia

Since the early nineties, information management researchers at the eScholarship Research Centre (ESRC) and its predecessor units at the University of Melbourne have been involved in exploring and utilising the capabilities of emerging digital and networking technologies in the provision of scholarly information infrastructure. The approach has been to identify the archival, library and scholarly principles embedded in traditional reference tools and then explore how they may be re-engineered and re-imagined with new information and communication technologies. This has led to a number of collaborative research projects with scholars and cultural institutions which involve building new digital information infrastructure respectful of diversity and complexity, and allow the exploration of new roles for various stakeholders in the processes to add richness, improve productivity and enable sustainability.

The latest such project involved working with the Australian Women's Archives Project on the redevelopment of the *Australian Women's Register* as collaborative information infrastructure. Technological development entailed:

- developing harvesting services by which content from the Register is made part of the National Library of Australia's exciting new *Trove* discovery service,¹ using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)² and the new Encoded Archival Context (EAC)³ metadata standard, and
- investigating the deployment of Web 2.0 tools into the system to allow for more efficient and effective content creation by community contributors.

This half-day workshop will use the redevelopment of the *Australian Women's Register* to explore issues around creating sustainable information infrastructure for the digital humanities. Questions around the themes of content, compliance, collaboration and complexity will be raised, illustrated with examples, and discussed with workshop participants. For example:

- **Content** – What new roles may the various stakeholders play creating and sustaining digital and networked information infrastructure? What impact does that have on the existing practices and systems of historical scholars, archivists, librarians and other information management professionals? What place does editorial and authority control have?
- **Compliance** – What kind of standards should the community look to influence and/or develop? What are the benefits of standards compliance? What are the costs?
- **Collaboration** – What new collaborations are made possible with the new technologies? What new dependencies? How are collaborations sustained? How are collaborative information networks made resilient and robust?
- **Complexity** – What information models support diversity and complexity? How can the development of open and interoperable systems be facilitated? What organisational and social factors mitigate their development?

Presenters

Joanne Evans is a Research Fellow at the University of Melbourne's eScholarship Research Centre and has been responsible for the design, development and deployment of the Centre's archival information systems in humanities and cultural heritage projects. With qualifications and experience in information management, recordkeeping and archiving, and systems development, her research interests lie in exploring ways in which library and archives principles are applied into scholarly practices in order to meet the challenges of the digital and networked age particularly for the humanities, arts and social sciences. Joanne has also been involved with recordkeeping and resource discovery metadata standards development as part of working groups within Standards Australia's IT 21/7 Committee and with the Australian Society of Archivist's Committee on Descriptive Standards.

Nikki Henningham is a Research Fellow in the School of Historical Studies at the University of Melbourne and is the Executive Officer for the Australian Women's Archives Project. She completed her PhD, a study of gender and race in Northern Australia during the colonial period, in the Department of History at the University of Melbourne in 2000. Since then,

she has taught in a wide range of undergraduate subjects, including world history, film and history and Australian history, and has conducted research for a variety of projects, including the Australian Women's Archives Project. She has research interests in the general area of Australian women's history, with a particular focus on women and sport, women and oral history and the relationship between the keeping of archives and the construction of history. In 2005, she received the National Archives of Australia's Ian Mclean Award for her work in this area.

Half day workshop: Morning, 6 July.

Notes

1. *Trove* is the National Library of Australia's new discovery service, providing a single point of access to resources held in Australia's memory institutions and incorporating rich contextual metadata from a variety of sources. See <http://trove.nla.gov.au/>.
2. OAI Protocol for Metadata Harvesting (OAI-PMH) is a lightweight harvesting protocol for sharing metadata between services developed by the Open Archives Initiative. It defines a mechanism for harvesting metadata records from repositories based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language) in support of new patterns for scholarly communication. See <http://www.openarchives.org/pmh/>.
3. Encoded Archival Context – Corporate bodies, Persons, and Families (EAC-CPF) is a metadata standard for the description of individuals, families and corporate bodies which create, preserve, use, are responsible for, or are otherwise associated with records. Its purpose is to standardize the encoding of descriptions of agents and their relationships to resources and to one another, to enable the sharing, discovery and display of this information. See <http://eac.staatsbibliothek-berlin.de/>.

Text Mining in the Digital Humanities

Heyer, Gerhard

gheyer@eaqua.net

eAQUA Project, Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany

Büchler, Marco

mbuechler@eaqua.net

eAQUA Project, Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany

Eckart, Thomas

teckart@eaqua.net

eAQUA Project, Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany

Schubert, Charlotte

schubert@eaqua.net

eAQUA Project, Ancient History Group, Department of History, Faculty for History, Art, and Oriental Studies, University of Leipzig, Germany

Thinking about text mining and its scope in the Digital Humanities, a comparison between the theory based work of the Humanities and the model driven approaches of Computer Science can highlight the decisive differences. Whilst Classicists primary rely on manual work e. g. using a search engine which just finds what is requested and skips non-requested but nonetheless interesting results, an objective model can be applied to the whole text and is closer to completeness. Even if the implication of the result doesn't depend on what the researcher does, the quality itself is typically worse than manual work. That's why the workshop combines both the quality of manual work and the objectivity of a model.

The workshop contains four sessions of 90 minutes as well as one hour for lunch (not provided) and two half-hour breaks (all in all 8 hours). Every session is segmented into three parts:

1. Theoretical background (30 minutes):

Within this section the necessary background is given to bring workshop relevant knowledge to the participants. This includes a soft brainstorming of the algorithms working behind the user interfaces.

2. Introduction of the user interface (15 to 30 minutes):

To avoid reading a manual a short introduction to the user interface is given. The short introduction of the presenter can be followed locally by every participant. When a problem

occurs, the non active presenters will help the respective participant.

3. Hands-on section (30-45 minutes): After receiving the text mining background and a short introduction to the user interface, the participants have up to half of a session for working on their own laptops. All presenters can be asked for detailed questions.

Based on the works within the eAQUA project of the last years, the modules *Explorative Search*, *Text Completion*, *Difference Analysis* as well as *Citation Detection* are chosen to highlight the benefits of computer based models. In detail that means:

- **Explorative Search:** By using Google in daily life almost everything can be found. The basic idea is: if a web page doesn't contain the information sought, any other will do. The differences in searching humanities texts can be grouped to two main clusters: a) The text corpus is closed and relatively small compared to the Internet; b) In relation to daily life queries on Google complete requests in the humanities are quite uncommon since the set of words are often unknown. For this reason a graph based approach is used to find (starting with a single word like a city or a person name) interesting associated words you would typically not have directly in your mind. At the end of this session, it will be discussed briefly how such an approach can be integrated into teaching since especially for students a search like this can be useful to explore and learn a domain.
- **Text Completion:** Because of the degree of fragmentation of papyri and inscriptions, a dedicated session for completing texts is set on the agenda. In this session well established approaches of spell checking will be combined with dedicated techniques addressing ancient text properties.
- **Difference Analysis:** In this session a web based tool is introduced to compare word lists of e.g. two authors, works or literary classifications. The result is divided into five categories: two categories containing words only used in one of the two text sets, two categories representing words which are significantly more often used in one of the two sets and finally a class of words with similar frequency. Based on these separations, differences can be identified faster than by manual reading.
- **Citation Detection:** The session of detecting citations contains three different aspects: a) How can citations be detected? b) How can found citations be accessed as efficiently as possible by Ancient Greek philologists (micro view on citations)?; c) How can more global associations be found like dependencies between centuries and dedicated passages of works (macro view on citations)? The main focus of this session is not set

on the algorithms to find citations but on both user interfaces for different research groups.

Full day workshop: Monday, 5 July.

Introduction to Text Analysis Using JiTR and Voyeur

Sinclair, Stéfan

sgs@mcmaster.ca

McMaster University, Canada

Rockwell, Geoffrey

geoffrey.rockwell@ualberta.ca

University of Alberta, Canada

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 JiTR collections management tool to do text analysis environment on electronic texts with Voyeur. JiTR (<http://ra.tapor.ualberta.ca/~jitr>), which stands for Just in Time Research is a platform for managing collections of texts which can launch text analysis tools like Voyeur. Voyeur is the latest text analysis web based system developed by TAPoR collaborators and it brings together visualization and concordance tools in a fashion that allows multipanel interactive analysis or single tool analysis. Voyeur is capable of scaling to handle multiple documents and larger texts than previous web based tools.

The workshop will provide

An introduction to managing texts with JiTR for analysis. This will include ways of aggregating texts for analysis and how to manage the tools you use in JiTR, especially Voyeur.

1. An introduction to different ways of using Voyeur. Voyeur can be used in a multi-panel view where the different views interact (see screen shot below) or as individual tools. Users will be shown different ways of running Voyeur and how to manage panels.
2. 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.
3. 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.

4. 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.

Participants are encouraged to bring their own documents on a Flash Drive for the workshop.

Half-day workshop: Morning 7 July.

Designing a Digital Humanities Lab

Veomett, Angela

veometta@berkeley.edu

University of California, Berkeley

Since 2008, the Townsend Center for the Humanities at UC Berkeley has been developing an online digital media lab to facilitate interdisciplinary and collaborative research projects in the humanities. The principle motivation for the project was to lead researchers in the humanities—students, faculty and the scholarly community-at-large—toward more creative methods of conducting research, and new ways of conceptualizing intellectual relationships through the use of Web 2.0 technologies. While many humanists have begun to use digital resources in their research, there is a segment of the community who, for various reasons, are not using even the most basic resources to their advantage. By providing these resources in a format familiar to academics, the Townsend Humanities Lab aims to be the place where humanities scholars can explore and experiment with digital resources.

In our workshop, we will address the challenges of developing and implementing such a project, including the institutional challenges that face small departments wishing to undertake a large project requiring ongoing support and the individual challenge of addressing the needs of traditional scholars in the humanities. For institutions interested in implementing a similar project, we will provide information about how we have supported and sustained this project, with open-source solutions possessing a key role in this process. For those interested in the broader issues of opening digital tools to a wider academic base, we will also discuss the importance of adapting functionality, interface design, and language to special constituencies.

In addition to these broader discussions, we will demonstrate our current implementation of the Townsend Humanities Lab, discussing the specific strategies used to attract and support humanities scholars. We will then guide workshop attendees through creating a Lab Project, demonstrate all of the resources available to project managers, and explain the social networking opportunities available to project members. We will also discuss future goals for the Lab, including integrating WorldCat and Zotero functionalities for organizing Project resources. Finally, as the Lab is in Beta phase, we wish to engage participants in a design critique

and a general discussion about the kinds of digital tools that are most helpful for interdisciplinary and collaborative projects.

Two hour workshop: Morning, 6 July

Peer Reviewing Digital Archives: the NINES model

Wheeler, Dana

dw6h@cms.virginia.edu
University of Virginia

Mandell, Laura

laura.mandell@gmail.com
Miami University of Ohio

The aim of this workshop is to invite digital humanists to work together in figuring out how to peer review digital scholarship. It springs from three impulses. The first has to do with the day-to-day work of NINES. Dana Wheeler is Project Manager and Laura Mandell Associate Director of NINES, the Networked Infrastructure for Nineteenth-century Electronic Scholarship (<http://www.nines.org>). Dana and Laura are engaged in helping scholars figure out how to create state-of-the-art scholarly editions, on a much smaller scale than King's Centre for Computing in the Humanities.

We would like to help people who have projects in early-to-mid stages of development learn about standards and best practices for their digital archives, editions, or artworks. Second, Dana Wheeler works with sites to develop a robust metadata system that makes possible their interoperability in the NINES universe. She will discuss metadata encoded in NINES RDF and demonstrate its practical values. Third, as Chair of the MLA Committee on Information Technology this year, Laura Mandell participated in an excellent workshop for department chairs and junior faculty, created by Susan Schreibman, about reviewing digital work for promotion and tenure. Organizations such as NINES can help junior scholars obtain the rewards they deserve for digital work through thoughtful peer review and documentation. We invite members of Promotion and Tenure committees as well as heads of humanities organizations to attend this workshop.

- 1st Part: Laura Mandell will present the MLA standards for Electronic scholarly editions as well as information from the Electronic Literature Organization. Best practices for sustainability culled from articles in the library sciences will also be presented, and then participants will spend some time working together looking at real projects in order to figure out how to implement this information practically. We will sketch out for each site examined workflow and interface design.
- 2nd Part: Dana Wheeler will demonstrate how to create the metadata needed for inclusion in

NINES. Though participants may be planning to submit their work to NINES, they need not be working in nineteenth-century studies for this demonstration to be valuable. For those who are developing sites or archives, learning what metadata is needed at early stages of development is a boon, and for those who wish to sponsor an organization like NINES, or incorporate NINES practices into an existing organization, our metadata and indexing systems provide an interesting model.

- 3rd Part: Finally, for the last third of the workshop, workshop leaders and participants will all work together in peer-reviewing digital archives and sites. In the process, we will sketch out the kinds of documentation needed for selecting and encouraging the best digital work as well as for creating in the space of interdisciplinary digital humanities a robust set of practices for supporting the professional development of the scholars who contribute. The draft we write at this workshop will go forward to be worked on during the NINES Summer Institute, 2011.

This workshop is directed at two audiences: producers of electronic scholarship and art, on the one hand, and arbiters of it, on the other. The latter are those who are actively engaged in creating and reforming the institutional structures in which these projects will live and thrive. They might be editors of journals, managers of specific institutional structures, directors of disciplinary organizations, librarians, bibliographers, chairs of departments, or interested collaborators.

Half day workshop: Afternoon, 6 July.

Panels

Digital Literacy for the Dumbest Generation - Digital Humanities Programs 2010

Clement, Tanya

tclement@umd.edu

Maryland Institute for Technology in the Humanities, University of Maryland, USA

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de

Institut für Deutsche Philologie, Universität Würzburg, Germany

McCarty, Willard

willard.mccarty@mccarty.org.uk

King's College London

Marc Bauerlein argues that undergraduates now and undergraduates to come soon are “the least curious and intellectual generation in national history.”¹ Dubbing them “the dumbest generation” and “mentally agile” but “culturally ignorant,” Bauerlein decrees that The Web hasn’t made them better writers and readers, sharper interpreters and more discerning critics, more knowledgeable citizens and tasteful consumers” (Bauerlein, *The Dumbest Generation* 110). The crux of this attack on digital culture lies in the link that Bauerlein and others (“Reading at Risk” xii) make between paper and digital texts: “the relationship,” Bauerlein explains, “between screens and books isn’t benign” (“Online Literacy is a Lesser Kind”). Like Bauerlein and the authors of the NEA report, Sven Birkerts maintains that book readers learn more because the book is a system that “evolved over centuries *in ways that map our collective endeavor to understand and express our world*” and that “the electronic book, on the other hand, represents—and furthers—a circuitry of instant access” (“Resisting the Kindle”). In contrast to this perspective, scholars and educators in the digital humanities have spent decades working with digital texts and arguing that advanced knowledge production is the primary function of using computational methodologies in the humanities (Busa 1980, 89; Smedt 2002, 90; McCarty 2005, 13).

The three papers in this panel will give an overview on university programs teaching digital humanities in the US, the UK and Germany. The first paper will treat undergraduate programs, the second graduate programs and the third will

describe in depth one PhD program. Like others before us (McCarty, Orlandi, Terras, Unsworth, “The Humanities Computing Curriculum”), we are especially interested in comparing these programs, because this allows us to consider a common understanding of the essential aspects of the work in digital humanities. On the other hand we are interested in analyzing the differences and to explore as much as possible the reasons for them. So an analytic charter of the curricula is complemented by a closer look at the institutional affiliations of the programs and the people mainly responsible for them. Undergraduate programs, for example, have to manage the challenge to offer an introduction not only into *digital* humanities but into the humanities in general while graduate programs have to determine what kind of knowledge they demand from the students entering them. Although our overall perspective on these programs is similar, not only the personality of the three authors but also the specific problems of the different forms of programs motivate quite different papers. Thinking about the work that scholars do in the digital humanities from the perspective of the work we need to do to produce culturally literate and critically savvy—that is, *intelligent*—students is essential.

Notes

1. Please see <http://www.dumbestgeneration.com/home.html>.

PAPER 1

An Undergraduate Perspective

Clement, Tanya

tclement@umd.edu

Maryland Institute for Technology in the Humanities, University of Maryland, USA

The scholarship done by the digital humanities community demonstrates that inquiry enabled by modes of research, dissemination, design, preservation, and communication that rely on algorithms, software, and/or the internet network for processing data deepen and advance knowledge in the humanities. Marc Bauerlein complains that undergraduates are passive consumers of “information,” that they convert history, philosophy, literature, civics, and fine art into information,” information that becomes, quite simply, “material to retrieve and pass along” (“Online Literacy”). In contrast, Wendell Piez and other digital humanities scholars insist that when we study “how digital

media are *encoded* (being symbolic constructs arranged to work within algorithmic, machine-mediated processes that are themselves a form of cultural production) and how they encode culture in words, colors, sounds, images, and instrumentation,” we are “far from having no more need for literacy;” in fact, the cultural work done by and through digital media requires that students “raise it to ever higher levels.”

At this time, however, discussion concerning undergraduate pedagogy within the digital humanities community remains limited and scattered. For instance, a search for the word “undergraduate” in the past five years of abstracts from the DH conference (or the joint ACH/ALLC conference) shows that there have been less than five presentations specifically concerning undergraduate pedagogy (Jessop 2005; Mahony 2008; Keating, *et al.* 2009). This trend may be linked to the notion that an undergraduate curriculum is more about teaching and less about research (Smedt, *et al.* 16), but this answer is reductive if not partially untrue. At the same time, if we believe that the work digital humanists do “can help us to be more humanistic than before” (Busa 89), why isn’t there more discussion within the DH conference and publications about this essential aspect of undergraduate study?

This paper will discuss work in place to lay the foundation for further (both in terms of more and deeper) discussion about undergraduate education in the digital humanities. First, this paper will present an updated and annotated bibliography of current undergraduate programs that are inflected by the digital humanities. Though Willard McCarty and Matthew Kirschenbaum’s list of “Institutional models for humanities computing” is extensive, it does not include an updated account of specifically undergraduate programs. That undergraduate studies are not well discussed within the DH community is part and parcel with the fact that it is a field that engages a wide range of disciplinary perspectives and it is a field that is represented by programs of study that are inflected by, but not necessarily called, Digital Humanities. Because this annotated bibliography will be developed as the result of an ongoing discussion with a disperse community, it will reflect a wide range of programs that the community has itself defined as “inflected by digital humanities.” Already, I have created an online list of undergraduate programs generated through an informal survey conducted on *Twitter*, the *Humanist Discussion List*, and the blog U+2E19.¹ To date, the website at King’s College London still touts itself as “one of the very few academic institutions in the world where the digital humanities may be pursued as part

of a degree” in undergraduate studies—a fact that is largely still true—but there are many programs without formal degrees where important pedagogy concerning digital culture happens.²

The fact that the list already includes a broad range of programs encompassing information science, digital cultures, new media, and computer science reflects the difficult nature of training an undergraduate student in the “methodological commons” (McCarty 131) of the digital humanities, but it also reflects the provocative nature of describing what that curriculum might look like. According to Unsworth, “the semantic web is our future, and it will require formal representations of the human record” requiring “training in the humanities, but also in elements of mathematics, logic, engineering, and computer science” (Unsworth). Patrik Svensson sees work in the digital humanities as a kind part of a spectrum “from textual analysis of medieval texts and establishment of metadata schemes to the production of alternative computer games and artistic readings of nanotechnology” (Svensson). Smedt and his colleagues choose to limit their definition of DH undergraduate programs in order “to concentrate on *computing* and to avoid the fields of information, communication, media, and multimedia since these are generally considered as social sciences rather than as humanities” (16). Just as asking the question “What is Humanities Computing and what is not?” (Unsworth) generates more questions, asking the community to identify programs inflected by the digital humanities is sure to provoke more discussion concerning existing models. What is important to teach these students? What is the core knowledge base needed?

When discussing current models, it is equally important to make transparent the institutional and infrastructural issues that are specific to certain universities. What works for one institution will not necessarily work for another. By the same token, simply providing examples of existing programs would belie the extent to which scholars and administrators shape these programs (whether they grant degrees, certificates, or nothing at all) according to the needs of their specific communities. Consequently, in order to make these matters transparent and broaden discussion about the broad range of issues that underpin the formation of an undergraduate curriculum, I will disseminate a survey to the digital humanities community asking basic questions concerning how an undergraduate program inflected by the digital humanities has been and might be developed within a variety of university settings. These questions are based on previous conversations (Hockey 2001; Unsworth, Butler 2001), but this previous work has focused primarily on graduate (or post-graduate) work. In

my attempt to update the conversation with a focus on undergraduate study, I incorporate questions that concern curriculum and questions, which point to infrastructural and institutional concerns that are specific to undergraduate education:

1. What are the aims and objectives of your undergraduate program?
2. How is the academic content of the program structured? What are the core modules/courses?
3. What are the academic backgrounds of students accepted for the program? Are there any particular requirements?
4. Does the program involve participation in research projects at area institutions or centers? If so, what factors influence which projects are chosen? How is participation monitored and assessed?
5. What is the program's relationship to the larger undergraduate community? Does the program include events, publications, or other opportunities for outreach? Does the program include a residential component, or other opportunities for community building?
6. Does the program grant a certificate or degree? What are the key issues in establishing a certificate or degree for students in your program?
7. How does the program fit into the overall structure of the institution?
8. Are there classes already being taught at your institution? What are the key issues in bringing these classes together under the rubric of a single curriculum?
9. What technical facilities are needed for the program and how are these supported?
10. What are other important infrastructural issues and challenges in setting up a program within your institution?

This paper will present and analyze the findings from this survey.

Finally, this paper will conclude with a case study describing the development of Digital Cultures and Creativity,³ an undergraduate living and learning program at the University of Maryland, College Park (UMD) that we have designed for the 21st century student who was born into the world of windows and the web. The result of a partnership between the Maryland Institute for Technology in the Humanities (MITH) and UMD's Arts and Humanities College, DCC is part of the university's new Honors College and will commence in the fall of 2010 with classes run by faculty from the Computer Science Department, the Information School, the Art Department and the

English Department. In an effort to make transparent how a program of this nature is developed across disciplines within a large research university, this paper will detail the various stages of development—curricular and administrative—we have navigated during the 2009-2010 planning phase.

In 2001, Steven Tötösy de Zepetnek observed that because undergraduates began their research online, scholars should create more and better online resources for academic study (Tötösy). A glance just at the last ten years of the journal of *Literary and Linguistic Computing*, the abstracts from the annual Digital Humanities conference, and the first issues of the *Digital Humanities Quarterly* prove that the DH community has worked hard to create these resources. Scholars in the digital humanities are already teaching the next generation of students not only how to use electronic resources, but how to create them, expand them, and preserve them. Now is the time to make that work transparent and to provide a resource for others who wish to continue, broaden, and support this work.

Notes

1. Please see Blog post "Digital Humanities Inflected Undergraduate Programs" at <http://www.palms.wordhers.net/wp/2009/11/digital-humanities-inflected-undergraduate-programs-2/>.
2. At the time of this writing, Martyn Jessop has written in the *Humanist* Discussion Group to clarify: "Sadly the . . . minor at King's College London has been closed down" though they "still operate 'standalone' modules in digital humanities for 1st and 2nd year students" (Jessop 2009).
3. Please see <http://dcc.umd.edu>.

PAPER 2

Graduate Programs

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de

Institut für Deutsche Philologie, Universität Würzburg, Germany

In the last five years new graduate programs for digital humanities have been developed in Germany or are actively developed at the moment in four to five universities. This could be understood as a sign that digital humanities have been accepted as part of a modern university and its spectrum of disciplines. On the other hand it may also point to the fact that in the context of the Bologna process in Europe (i.e. recreating all programs in BA and MA formats to allow students more freedom to change their place of study) Digital Humanities is seen by many people

in university administration as uncommon enough to provide the possibility of contributing effectively to the profile of a university. For some, Digital Humanities seems also to contribute a solution to the problem or the tension between the self image of German universities as institutions which are not tasked with providing practical knowledge immediately useful for any profession, on the one hand, and the demands of students and the public in general at least to diminish the gap between the knowledge taught at the university and the demands of professional life outside it.

This paper will give an overview of these new graduate programs in Digital Humanities in Germany and will compare them to those in the US and the UK. As some programs are in the process of being developed just now, there are no definite results at the moment. The analysis will follow these questions:

- What previous knowledge is required by the programs, especially, in respect to computer science, technical skills and the humanities?
- How is the knowledge taught by the programs modeled: primarily as practical, as conceptual, as theoretical, and what is the dominant knowledge model for the program: for example computer science, some form of applied science, humanities?
- How is the program positioned institutionally: What school or department does it and the larger part of its teaching staff belong to? What kind of degree can students earn with the program? Can the program be combined with others and from what kind of department?
- How concrete or abstract is the program defined? Is there a detailed course description or a syllabus of required texts?
- What elements of the program can be found in more general humanities, library science or computer science programs?

There will be a systematic chart comparing the German programs with a selected list of programs in the US and the UK based on the aspects mentioned above and a systematic analysis of earlier research on digital humanities education. But the paper will also include anecdotal evidence on the problems and difficulties of developing and maintaining such a program.

It will be of special interest to find communalities between all these programs. It is to be expected to find some on the level of practical knowledge like text encoding using XML and TEI, but it is an open question whether more abstract competences like data modeling are shared by all or even most. The same is true for knowledge and expertise usually associated with the humanities: the ability to analyze

fictional texts and art, in this case digital texts and digital art.

It has been pointed out that in the long run the discussion on curricula is probably more important for our conception of digital humanities than the theoretical discussion about the topic or at least it is an important contribution to the discussion (cf. Terras). This paper also understands each curriculum as a statement in this discussion, a statement not only directed at the digital humanities community but also towards the humanities. This statement delineates what kind of knowledge about the digital is taken for granted as part of the humanities and what is (still?) marked as special knowledge not shared by most humanities scholars. This borderline is also under discussion and an analysis of what kind of modules and courses are taught by the staff mainly engaged in the digital humanities program can show what point these negotiations have reached.

At last some questions will be discussed which cannot be answered based on the information available now: For example, what kind of professional profile will students with degrees in those programs have? Will digital humanities as a discipline recruit its academic teachers in future from these new programs? Even if the answers cannot be given at the moment, maybe the discussion can contribute to the reflection on question of what a successful graduate program in the digital humanities would look like.

PAPER 3

PhD in Digital Humanities, King's College London

McCarty, Willard

willard.mccarty@mccarty.org.uk

King's College London

1. The PhD in the UK and at King's

In the United Kingdom the PhD (DPhil at Oxford) is normally a 3-4 year, research-only degree (6-8 years part-time). No course-work or qualifying examinations are required. At King's College London the candidate enters officially into an MPhil; after 9 months to a year he or she then applies for an upgrade to PhD. To gain this upgrade the candidate must demonstrate that he or she is producing work to the standards of a PhD.

2. The PhD in Digital Humanities at King's

The Centre for Computing in the Humanities, King's College London, has offered the PhD in Digital Humanities since 2005. Its first student dropped out after approximately a year, but since then the programme has taken on a further ten. Its first graduate will likely take his degree in Autumn 2010. All of the students have come to the PhD from other institutions, 3 from the U.K., 5 from elsewhere in the European Union, 1 from Norway and 1 from the United States. Of the total 3 are enrolled part-time, the remainder full-time. None has had prior degree-training in the subject. Currently 3 others are in process of developing their research proposals before applying. Since the programme was created there have been in addition 41 serious enquiries, likely at least a dozen of which would have proceeded to enrolment had adequate funding been available.

Apart from two studentships, one from the School of Humanities in 2006 and another from the Arts and Humanities Research Council in 2009, the programme has had no funding. Teaching assistantships are not available, but the department has begun to offer limited part-time research positions for work on collaborative projects. Lack of funding remains the most serious impediment to growth of the programme. No serious efforts have yet been made to advertise it.

The primary criterion for admission to the PhD in Digital Humanities is a cogent research proposal supported by letters of recommendation. Usually the applicant develops this proposal in consultation with the departmental Director of Research (Professor McCarty). Proposals of sufficient quality are accepted providing that the department, possibly in collaboration with one or more other departments, can support the research adequately. In a majority of cases to date (7 out of 10) supervision is cross-departmental: 3 with History, 1 with German, 1 with Portuguese, 1 with Computer Science, 1 with the Department of Education and Professional Studies, School of Social Science and Public Policy. Of these 2 are associated also with the King's cross-school Centre for Language, Discourse and Communication. Usually the balance of supervision is equally divided between the CCH and the other department but can vary from 70% to 30%.

Of the current students, 8 out of 10 have come to the programme directly rather than on referral from other departments. In other words, the PhD in Digital Humanities is the primary attractor for those wishing to involve computing as a major component in their research.

From 2009 students for whom living in London would pose a significant hardship can with approval pursue a "semi-distance PhD", with face-to-face supervisory meetings according to an agreed schedule and meetings by Skype as needed. Two students now take advantage of this arrangement, one full-time, one part-time.

Apart from individual supervision, all students in the programme meet in the monthly PhD Seminar, face-to-face or via the Internet, to present and discuss their work. Some meetings are partly devoted to special topics of interest to all. The PhD Seminar also includes students at the University of Alberta and, in a special credit-course, students at the University of Victoria, British Columbia, in real-time via Skype and Dimdim.

3. Development of the PhD in Digital Humanities

The PhD in Digital Humanities has been shaped primarily by the interests of applicants rather than by a pre-conceived notion of what a doctoral degree in the field should be. Technical competence at a level appropriate to computer science or information science has not been assumed or required, although critical thought on computing has been stressed from the beginning through development of research proposals and required subsequently. Practical work is strongly encouraged though it has not been required. A central chapter on the relevant computing methods has become the norm, with a thorough knowledge of the secondary literature attested in a survey or in the citations. In a few cases students have needed and been given technical support from within the department to develop an application of computing. In equally many cases, however, students have come with a high level of technical knowledge and skills. In two cases arrangements are being made to provide specialised training, and in one the student separately obtained an 18-month paid fellowship to work abroad in a technical research institution (some of this work with engineers to design and construct relevant hardware). In brief, highest priority has been given to critical reasoning with and about computing in a manner consistent with the interpretative, problematizing disciplines of the humanities.

In all cases the subject of the research has been a problem within or recognizable to one or more of these disciplines. We have assumed that those who wish to practice computer science on material usually studied in the humanities are altogether better served by that discipline, although we are open to requests for collaborative supervision originating in computer science. In general no decision has been made *a priori* to restrict primary areas of investigation to

the humanities. Students from elsewhere are most welcome to apply, especially since they may well assist us in extending an already broad church.

References

- Bauerlein, Mark** (2008). *The Dumbest Generation: How the Digital Age Stupefies Young Americans and Jeopardizes Our Future (or, Don't Trust Anyone Under 30)*. New York, NY: Jeremy P. Tarcher/Penguin, Print.
- Bauerlein, Mark.** 'Introduction'. *The Dumbest Generation*. Web.
- Bauerlein, Mark.** 'Online Literacy Is a Lesser Kind'. *The Chronicle of Higher Education*. 19 September 2008, Web (accessed 11 Nov 2009).
- Birkerts, Sven.** 'Resisting the Kindle'. *Atlantic Monthly*. 2 March 2009, Web (accessed 15 Nov 2009).
- Burnard, Lou** (1999). *Is Humanities Computing an Academic Discipline? or, Why Humanities Computing Matters*. Web (accessed 11 Nov 2009).
- Busa, Roberto** (1980). 'The Annals of Humanities Computing: The Index Thomisticus'. *Computers and the Humanities*. 14: 83-90, Print.
- Hockey, Susan** (2001). *MA Programmes for Humanities Computing and Digital Media*. New York University, Web (accessed 11 Nov 2009).
- The Humanities Computing Curriculum: The Computing Curriculum in the Arts and Humanities*. Malaspina University College, Nanaimo, British Columbia, Canada, November 9-10, 2001, Web (accessed 15 Nov 2009).
- Jessop, Martyn.** 'In Search of Humanities Computing in Teaching, Learning and Research.'. *The International Conference on Humanities Computing and Digital Scholarship, The 17th Joint International Conference: Conference Abstracts (2nd Edition)*. Pp. 91-93, Web (accessed 15 Nov 2009).
- 'undergrad programmes'. *Humanist Discussion Group*. 27 Oct 2009, Web (accessed 11 Nov 2009).
- Keating, John J., Teehan, Aja, Byrne, Thomas** (2009). 'Delivering a Humanities Computing Module at Undergraduate Level: A Case Study'. *Digital Humanities 2009 Conference Abstracts*. Pp. 167-169, Web (accessed 15 Nov 2009).
- Mahony, Simon** (2008). 'An Interdisciplinary Perspective on Building Learning Communities Within the Digital Humanities'. *Digital Humanities 2008 Conference Abstracts*. Pp. 149-151, Web (accessed 15 Nov 2009).
- Mangen, Anne.** 'Hypertext Fiction Reading: Haptics and Immersion'. *Journal of Research in Reading*. 31.4.2008, Print.
- McCarty, Willard.** 'New Splashings In The Old Pond: The Cohesibility Of Humanities Computing'. *The Alliance of Digital Humanities Organizations*. 16 October 2002, Web (accessed 15 Nov 2009).
- McCarty, Willard, Kirschenbaum, Matthew.** 'Institutional models for humanities computing'. 12 Dec 2003, Web (accessed 15 Nov 2009).
- Nowviskie, B., Unsworth, J. (eds.)** (1999). *Is humanities computing an academic discipline? An interdisciplinary seminar*. University of Virginia, Autumn 1999, Print.
- Piez, Wendell** (2008). 'Something Called "Digital Humanities"'. *Digital Humanities Quarterly*. 2.1, Web (accessed 11 Nov 2009).
- Orlandi, Tito.** 'Is Humanities Computing a Discipline?'. *The Alliance of Digital Humanities Organizations*. 17 May 2002Web (accessed 11 Nov 2009).
- Unsworth, J., Butler, T.** (2001). 'A Masters Degree in Digital Humanities at the University of Virginia'. *Session, ACH-ALLC 2001*. New York University, June 13-16, 2001 11 Nov 2009, Web.
- de Smedt, Koenraad** (2001). 'Some Reflections on Studies in Humanities Computing'. *Literary and Linguistic Computing*. 17.1: 89-101, Print.
- Smedt, Konrad, Gardiner, Hazel, Ore, Espen, Orlandi, Tito, Short, Harold, Souilliot, Jacques, Vaughan, William (eds.)** (1999). *Computing in Humanities Education: A European Perspective*. Bergen: University of Bergen, Print.
- Svensson, Patrik** (2009). 'Humanities Computing as Digital Humanities'. *Digital Humanities Quarterly*. 3.3, Web.
- Terras, Melissa** (2005). 'Disciplined: Using Curriculum Studies to Define 'Humanities Computing''. *Abstracts ACH/ALLC 2005*. Web (accessed 15 Nov 2009).
- Tötösy de Zepetnek, Steven** (2001). 'The New Knowledge Management and Online Research and Publishing in the Humanities'. *CLCWeb: Comparative Literature and Culture*. 3.1, Print.
- Unsworth, John** (2002). 'What is Humanities Computing and What is not?'. *The Alliance of Digital Humanities Organizations*. Web (accessed 11 Nov 2009).

Computational approaches to textual variation in medieval literature

van Dalen-Oskam, Karina

karina.van.dalen@huygensinstituut.knaw.nl

Huygens Instituut, Netherlands

Thaisen, Jacob

thaisen@ifa.amu.edu.pl

Adam Mickiewicz University, Poznań, Poland

Kestemont, Mike

mike.kestemont@gmail.com

CLIPS Computational Linguistics group and
Institute for the Study of Literature in the Low
Countries (ISLN), University of Antwerp, Belgium

Before the age of printing, texts were copied manually only. This was done by scribes – persons who made a copy of a text for their own use or for the use of others. Often, the original is no longer available. All that remain are copies of the text, or copies of copies of copies. We know that scribes made mistakes, and that they changed spellings and wording according to what they thought fit for their audience. And we know that they sometimes reworked the text or parts thereof.

Up till now, these insecurities may have made medieval texts less interesting to work on for digital humanists. However, the complex world of medieval textual copying is a very challenging topic in its own right. Recently, some scholars have tried to develop and apply digital methods and techniques to gain insight in manual text transmission. In this session, they will explain which specific research questions led to their approach, and why traditional methods did not suffice. Then they will describe the digital approach they developed, how they gathered their data, and present the first results. They will sketch the next steps for their research and reflect on which larger questions may come closer to an answer, and which other areas of digital humanities will benefit from this research.

The first paper (by Jacob Thaisen) will focus on how the variability of spelling characteristic of Middle English makes probabilistic models a powerful tool for distinguishing scribes and exemplars. The second paper (by Karina van Dalen-Oskam) goes into vocabulary and frequencies of parts of speech, as a means to get insight in the influence scribes exerted on the texts they copied. The third paper (by Mike Kestemont) aims at erasing or minimizing

textual differences in order to assess stability and the persistence of authorial features of manually copied medieval texts.

PAPER 1

Probabilistic Modeling of Middle English Scribal Orthographies

Jacob Thaisen

thaisen@ifa.amu.edu.pl

Adam Mickiewicz University, Poznań, Poland

With the Norman Conquest of 1066 written English ceased to be employed for administrative and other official purposes, and the normative spelling conventions established for the West Saxon variety of Old English fell into disuse. When the language eventually regained these crucial domains around three centuries later, a norm for how to spell English no longer existed. The only models available to scribes were the practices of other languages known to them or, increasingly as English strengthened its position, the conventions adopted in the exemplars from which they copied. As a result of the interaction of all these factors, Middle English—the English of the period from the Battle of Hastings to Caxton's introduction of printing from movable type in 1476—is characterized by considerable variation in spelling, even within the output of a single individual. There is nothing at all unusual about one and the same scribe of this period representing one and the same word in more than one way, including very frequent words such as the definite article and conjunctions. Moreover, scribes could use the variability to their advantage in carrying out the copying task, for example, to adjust the length of lines or speed up the copying process.

The variability of Middle English orthography means it would be misguided to assume that two texts penned by a single scribe necessarily follow, or should follow, identical spelling conventions. They are much more likely to exhibit variation within bounds. Any stylometric attribution of Middle English texts to a single scribe or of portions of a text in a single scribal hand to different exemplars on the basis of spelling must take this nature of the evidence into account. The probabilistic methods known from statistically-based machine translation, spell-checking, optical character recognition, and other natural language processing applications are specifically designed to recognize patterns in "messy"

data and generalize on the basis of them. It is the purpose of this paper to demonstrate that this property of these methods makes them adequate stylometric discriminators of unique orthographies.

The methodologies developed in connection with the preparation of *A Linguistic Atlas of Late Medieval English* (McIntosh, Samuels, et al. 1986) separate unique orthographies by manual and predominantly qualitative means; if quantitative data are collected at all, they are subjected only to simple statistical tests. Since texts differ lexically, they are not readily comparable in all respects. The *Atlas* solution is to generate comparability by restricting the investigation to the subset of the respective lexicons of the various texts they may reasonably be expected to share, such as function words. Spelling forms for these words are collected from samples of the texts by selective questionnaire and any pattern present in their distribution detected by visual inspection. The forms are further often analyzed by reference to known dialect markers. The latter translates as the researcher relating the forms to phonological and morphological variables, although there is recognition in the dialectological literature that geographic significance too may characterize other levels of language.

However, it is now practically feasible to estimate the full orthography of which a given text is a sample by building probabilistic models. The reason is that recent years have witnessed an increase in the amount of diplomatically transcribed manuscript materials available in digital form, which makes it possible to abandon the qualitative focus. Scholars are already subjecting the lexical variation present in similar materials to sophisticated computer-assisted quantitative analysis (Robinson 1997, van Dalen-Oskam and van Zundert 2007). Their studies point the way forward.

The building blocks of Middle English orthographies are not individual letters but sequences of letters of varying length which, further, combine with one another in specific ways, with phonograms, morphograms, and logograms existing side by side. Every Middle English orthography has a slightly different set of building blocks, making n -gram models a good type of probabilistic model for capturing the distinct properties of each. Such a model is simply an exhaustive listing of grams (letters and letter sequences), each with its own probability and weight.

"Perplexity" expresses how well a given model is able to account for the grams found in a text other than the one from which the model itself is derived. That is, a model – itself a list of grams – is compared with a list of the grams found in another text and the measure simply expresses the level of agreement

between the two lists. However, to find out whether the two texts are instances of the same orthography, a better model is a model not of the text from which it is derived but of the orthography which that text is a sample of. This is because the lexis of the text means the probabilities of the grams are not those they have in the orthography. This skew can be reduced by generalizing the model. "Smoothing" refers to the act of (automatically) introducing weights to achieve the best possible generalization.

Chen and Goodman (1998) carry out a systematic investigation of how a range of smoothing techniques perform relative to one another on a variety of corpus sizes in terms of the ability to account for test data. Their data come from present-day English and their basic unit is the word rather than, as here, the letter. They find the technique developed by Witten and Bell (1991) consistently to generalize the least effectively, and that developed by Kneser and Ney (1995), and later modified by Chen and Goodman (1998), consistently to do so the most effectively. Both weight every $(n-1)$ -gram in proportion to the number of different n -grams in which it occurs in the training data, i.e. in the text on which the model is based in the present case. The former technique produces the effect that the probability mass is shifted toward those grams which best characterize the training data, making it appropriate if the purpose is to distinguish orthographies within the product of a single scribe. The latter does the opposite, thus more fully capturing the full range of forms accepted by the scribe of the training data; this makes it the better choice if the purpose is to compare texts by a range of scribes in terms of their orthographic similarity.

To demonstrate the adequacy of smoothed models as discriminators of Middle English orthographies, the presenter investigates two corpora by means of the *SRI Language Modeling Toolkit* (Stolcke 2002):

1. The copy of Geoffrey Chaucer's unfinished poem *Canterbury Tales* contained in Cambridge University Library, MS Gg.4.27 [Gg]; the copy is in a single scribal hand.

The Gg text of is divided into equal-sized segments, each of which is subsequently modeled (Witten-Bell smoothing). For every model, its perplexity is computed against every segment other than the segment on which it is based, giving a 19x19 matrix with one blank cell per row. The mean and standard deviation is calculated for each row.

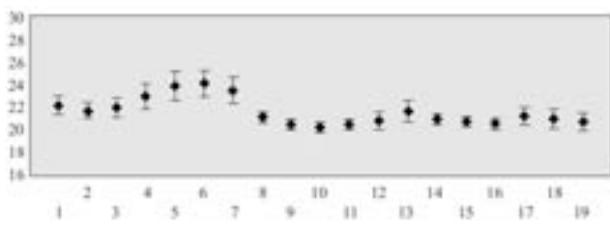


Figure 1

As can be seen in Figure 1, the greatest change in mean perplexity between any two consecutive segments falls between sections 7 and 8, with only that change falling outside the confidence intervals indicated by the whiskers. The hypothesis that two distinct populations are present is confirmed statistically (Mann-Whitney, $U = 84$, $n_1 = 7$, $n_2 = 12$, $P < 0.001$, two-tailed): thus, Gg contains two orthographies, their boundary falling around segments 7 and 8. The manuscript contains physical evidence of a change of exemplar late in segment 7 (Thaisen forthcoming).

2. 58 pre-1500 manuscript copies of the *Miller's Tale* and all fifty-eight such copies of the *Wife of Bath's Prologue*, totaling 116 texts; a range of scribes are responsible for these copies.

The Toolkit builds a model of every text and smoothes them (Kneser-Ney modified). For every model, its perplexity is calculated with respect to every text. The perplexities are arranged into two matrices for hierarchical clustering, one for the Miller-based models and another for the Wife-based models.

It is found, firstly, that the two trees are virtually mirror images of one another; secondly, that a Miller text and a Wife text which come from the same manuscript usually appear as sisters and that the cases in which they fail to do so are attributable to a change of scribe or exemplar posited on outside evidence (Thaisen 2009).

These results are sufficiently encouraging to warrant the investment of further resources. They show that probabilistic modeling offers a repeatable, quantified means of measuring the level of similarity between Middle English orthographies and so, also, a tool for separating them. That separation is important not only in authorship attribution and textual criticism, but also in manuscript studies and English historical linguistics. Additional advantages over the *Atlas* methodologies, which focus on dialect rather than textual studies, are the level of exhaustiveness, since all the available data are considered, as well as simple ease, the input being an unlemmatized transcript in plain text format.

References

- Chen, S. F. and Goodman, J. T.** (1998). *An Empirical Study of Smoothing Techniques for Language Modeling*. Technical report TR-10-98. Harvard University. <http://research.microsoft.com/en-us/people/joshuago/publications.htm> (accessed 11 March 2010).
- Kneser, R. and Ney, H.** (1995). 'Improved Backing-off for m-Gram Language Modeling'. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. , pp. 181-84.
- McIntosh, A., Samuels, M. and Benskin, M.** (1986). *A Linguistic Atlas of Late Mediaeval English*. Aberdeen: Aberdeen University Press.
- Robinson, P.** (1997). 'A Stemmatic Analysis of the Fifteenth-Century Witnesses to the Wife of Bath's Prologue'. *The Canterbury Tales Project Occasional Papers: Vol. II*. Blake, N. F. and Robinson, P. (ed.). London: Office for Humanities Communication, pp. 69–132.
- Stolcke, A.** (2002). 'SRILM: An Extensible Language Modeling Toolkit'. *Proceedings of the 7th International Conference on Spoken Language Processing*. Hansen, J. and Pellom, B. (ed.). Denver: Casual Productions, pp. 901–04.
- Thaisen, J.** (2009). 'Statistical Comparison of Middle English Texts: An Interim Report'. *Kwartalnik Neofilologiczny*. **56**: 205–21.
- Thaisen, J.** (forthcoming). 'A Probabilistic Analysis of a Middle English Text'. *Digitizing Medieval and Early Modern Material Culture*. Nelson, B. and Terras, M. (ed.). Tempe: Arizona Center for Medieval and Renaissance Studies.
- Van Dalen-Oskam, K. and van Zundert, J.** (2007). 'Delta for Middle Dutch: Author and Copyist Distinction in Walewin'. *Literary and Linguistic Computing*. **22**: 345–62.
- Witten, I. H. and Bell, T. C.** (1991). 'The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression'. *IEEE Transactions on Information Theory*. **37**: 1085–94.

PAPER 2**Distinguishing medieval authors and scribes****Karina van Dalen-Oskam**

karina.van.dalen@huygensinstituut.knaw.nl

Huygens Instituut, Netherlands

We know that medieval scribes (women or men manually copying texts) changed the texts they were copying. Scribes not only made mistakes, but also deliberately changed a text's spelling and wording. We also know that they sometimes changed the content of a text, leaving out episodes and adding others, or for instance changing the moral message. In these cases, scholars may want to describe a text not as a copy of another text but as an adaptation of it. It is not exactly clear, though, when to call a text a copy or an adaptation, and how often scribes chose to adapt instead of 'just' copy a text. It is also uncertain if all copies/adaptations of a medieval text survived and how many were lost (and why). To make things even more complicated, the exact date of most medieval manuscripts is uncertain. And in many cases we do not know the identity of the author of the original text or of the scribes.

1. Research questions

One of the tasks of scholars of medieval literature is to analyse the adaptations in a copy and to try to explain them in a poetical, ethical, or political context, which is of course difficult if the original version of a text is not extant. However, comparing different copies/adaptations of the same text usually presents scholars with enough data to make relevant observations and draw at least some conclusions. Until now, the depth of the analyses was limited to what the human eye and a scholar's amount of research time allowed. However, digital texts and digital text analysis tools can help us to compare texts in many more aspects than was possible up till now.

The questions that are of interest to us are: can we compare manual copies of the same text (semi-)automatically and get insight into the divergences which occur? Can we filter out differences that have to do with language development? Can we filter out the influence of subsequent scribes of a text and focus on those aspects which show us the original author most clearly? If so, could we apply (adapted versions of) authorship attribution tools to medieval texts? Could we also distinguish scribes from each other and are they distinguishable in the same or in a different

way from how authors can be distinguished from each other? Can we develop new tools or fine-tune existing tools for scribal measurements? And can these measurements decide if a text is a copy or an adaptation and if so, how radical the adaptation is?

Up till now, scholars hardly ever tried to systematically answer these questions. The necessary amount of work seemed not proportionate to the possible results as long as there still was enough low-hanging fruit in the close-reading type of analysis of text adaptation. Possibly, scholars in the course of time have been trained to NOT ask these impossible-to-answer questions, although two topics have always had a special place in the humanities: building family trees of manuscripts (stemmatology) and authorship attribution based on traditional, close-reading and simple statistical methods. This shows there has always been a keen interest in new and complex methods when they could possibly answer pressing questions.

2. Data

We would like to introduce two methods which may help scholars to gain insight in the amount of differences between copies of the same text. For this research, we are not interested in mere spelling differences but in more content-related differences. Our area of research is Middle Dutch literature. The first method is a rather simple approach to the vocabulary of all the copies, for which we needed lemmatization of the data, and the second is the comparison of part of speech frequencies in the copies of the same text, which implied a dataset tagged for Part of Speech. A corpus answering these needs was not available yet, so we had to create one first.

Not many Middle Dutch texts are extant in a substantial number of copies. We chose a work by the Flemish author Jacob van Maerlant: the *Rijmbijbel* ('Rhyming Bible'), which is a translation/adaptation of the Medieval Latin *Historia scholastica* written by Petrus Comestor. Van Maerlant finished this work in 1271, and many fragments and fifteen near-complete manuscripts (though not all containing all parts of the text) survive, dating from ca. 1285 to the end of the fifteenth century. One of these manuscripts is available in a good edition; it is also digitally available lemmatized and tagged for parts of speech. Transcriptions of the other manuscripts had to be made for this research. Because of the length of the texts (almost 35,000 lines), we had to work with samples. We chose 5 samples of 200-240 lines from different parts of the text, and transcribed the parallel texts (if available) from all 15 manuscripts, lemmatized the samples and tagged them for parts of speech. The manuscripts are indicated by the

letters A, B, C, D, E, F, G, H, I, J, K, L, M, N and O. The lemmas we added to the transcribed texts have the form of the Modern Dutch dictionary entry (or the form the Modern Dutch entry would have had, had the word survived into present-day Dutch). We differentiated between ten parts of speech: noun, proper name, adjective, main verb, copula / auxiliary verb, numeral, pronouns, adverb, preposition, conjunction.

3. Methods

We approached the samples as 'bags of words'. We made use of perl scripts listing all lemmas and parts of speech for each small sample and for the frequency measurements. For each part of speech, in each sample we measured the absolute frequency, the relative frequency, the average of the fifteen samples, the standard deviation, the z-score and the ranking of the manuscript in comparison with the other fourteen manuscripts.

4. First results of the comparison of vocabularies

Table 1 below lists the amount of words (lemmas) each text episode has which do not occur in any of the other copies of the same episode: unique words. The manuscripts A - O are listed in chronological order (although many dates are very approximate). The order from left to right agrees with the order of the episodes in the text itself.

Unique	Eva 'E'	Debora 'D'	Judit 'F'	NT 'M'	Josephus T'	Total
C	7	4	2	0	10	23
B	0	0	2	2	3	9
M	3	5	0	3	1	12
A	5	3	1	1	2	12
G	7	6	6	11	8	38
D	0	1	2	3	5	9
L	0	0	0	1	1	2
K	4	5	1	2	2	14
F	8	26	24	10	57	125
P	5	4	3	6	1	19
J	9	7	4	2	5	27
N	3	5	6	8	2	24
H	13	5	9	5	18.8	32
O	2	5	7	8.8	8.8	14
I	11	28	33	8.8	8.8	118
Total	79	104	154	54	95	516

Table 1

At a glance we can see that manuscripts E and I show the most unique words. This needs to be investigated: are these scribes the most radical in their adaptations? We will try to push this use of vocabulary analysis further, e.g. measuring the percentage of overlapping vocabulary in the episodes in the different manuscripts.

5. First results of the comparison of PoS frequencies

Figure 2 shows the relative frequencies of nouns for each of the five samples (E, D, J, M, T) for each of the

manuscripts (A - O). For this part of speech, we see a big change in the trend for the episode in all of the manuscripts for episode T in manuscript E and for episode J in manuscript I. Graphs for other parts of speech show the same trend.

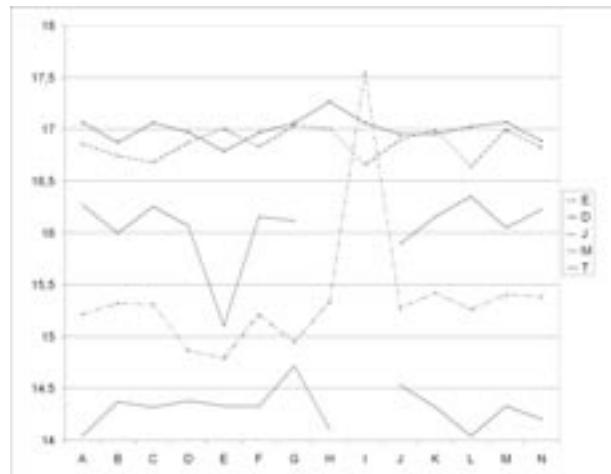


Figure 2

6. Evaluation

Reading the two episodes and comparing them with the other copies of the same text reveals that episode J in manuscript I clearly is a new text. Almost all the rhyme words have been changed, text is added, wording is completely different. The scribe here acted as a new author. This is not the case for episode T in manuscript E. Some rhyme words have changed, some wording is different, but the text is still clearly recognizable as a copy.

7. Conclusions and next steps

It seems simple comparisons of vocabulary and frequencies of parts of speech can pinpoint scribes who did more than copying their exemplar. If this can be confirmed by other experiments, this approach could help to direct scholars to those episodes in texts that may be most rewarding for a closer analysis (e.g. with traditional methods such as close reading). Before this is possible, however, a lot more medieval texts need to be available tagged for headwords and for part of speech. For that, a good tagger for Middle Dutch is highly desirable.

The research questions addressed above are key for getting a better insight into the cultural role of texts and the persons responsible for texts and their transmission, not only in the Middle Ages, but also later. It could help us to find a way to less subjectively compare texts and describe scribal adaptations, and in this way yield insight in the possible kinds of text manipulation throughout the ages.

References

- Dalen-Oskam, K. van and Zundert, J. van** (2008). 'The Quest for Uniqueness: Author and Copyist Distinction in Middle Dutch Arthurian Romances based on Computer-assisted Lexicon Analysis'. *Yesterday's words: contemporary, current and future lexicography. [Proceedings of the Third International Conference on Historical Lexicography and Lexicology (ICHLL), 21-23 June 2006, Leiden]*. Mooijaart, M., van der Wal, M. (eds.). Cambridge: Cambridge Scholars Publishing, pp. 292-304.
- Dalen-Oskam, K. van and Zundert, J. van** (2007). 'Delta for Middle Dutch – Author and Copyist Distinction in Walewein'. *Literary and Linguistic Computing*. **22**: 345-362.
- Kestemont, M. and Van Dalen-Oskam, K.** (2009). 'Predicting the Past: Memory-Based Copyist and Author Discrimination in Medieval Epics'. *Proceedings of the Twenty-first Benelux Conference on Artificial Intelligence (BNAIC 2009)*. Eindhoven, 2009, pp. 121-128.
- Spencer, M. and Howe, C. J.** (2001). 'Estimating distances between manuscripts based on copying errors'. *Literary and Linguistic Computing*. **16**: 467-484.
- Spencer, M. and Howe, C. J.** (2002). 'How accurate were scribes? A mathematical model'. *Literary and Linguistic Computing*. **17**: 311-322.

PAPER 3

The Robustness of Rhyme Words in Bypassing Scribal Variation for Medieval Authorship Verification

Mike Kestemont

mike.kestemont@gmail.com
CLIPS Computational Linguistics group and
Institute for the Study of Literature in the Low
Countries (ISLN), University of Antwerp, Belgium

1. A problem

Modern stylometric approaches can discriminate between authors to a fairly accurate extent. Machine learning techniques, for instance, are able to

'recognize' authors – be they literary or not – based on linguistic features extracted from representative textual samples written by these authors. In recent years, computational studies into authorship issues (such as verification and attribution) have been a popular topic in many research areas. Nevertheless, this kind of research has paid rather little attention to medieval literature. This is especially remarkable since it is precisely in this branch of philology that scholars have to contend with large amounts of texts of which the authorship is unknown or at least disputed. This lack of interest in medieval literature seems due to a variety of factors that all come down to the same basic fact: medieval texts are difficult to automatically process. For example, tools to perform basic actions such as the automatic lemmatization of texts are virtually non-existent for medieval languages, while most stylometric approaches heavily rely on e.g. lemma-frequencies for their feature extraction (cf. Burrows's *Delta*). This is mainly due to the scribal variation that is so typical of medieval manuscripts, as put forward in the introduction to this session proposal. Moreover, medieval texts are rarely extant from autographs and as such, in the majority of the cases, scholars have a hard time assessing which features in manuscript copies are *authorial* rather than *scribal*.

2. A solution

Any solution to the problem of authorship attribution for medieval texts has to overcome the difficulties imposed by the (scribal) instability of text in the Middle Ages. Whereas the other two papers in this session focus on the *exploitation* of scribal variation (i.e. textual instability), this paper aims at the exact opposite: *erasing* or *minimizing* these differences in order to assess textual stability and the persistence of authorial features of manually copied medieval texts. In this paper we shall focus on two methods to achieve this goal. Firstly, we shall briefly discuss the MiDL-architecture, a Natural Language Processing system designed for the automated tokenization, lemmatization and part-of-speech tagging of Middle Dutch literary texts. The techniques allow us to get past or 'transcend' superficial scribal variation and focus on the underlying authorial features of texts.

Secondly, we shall report on experiments with rhyme words and pairs – Middle Dutch epic literature was rhymed in about 99% percent of the cases that are currently known to us. This category of words is often claimed to be a very stable factor during the process of text transmission and thus can be expected to be extremely revealing with regard to authorial style. Scribes could not easily alter the rhyme words or rhyme scheme of a text without having to adapt

several lines of the text and would often refrain from doing this.

3. A good case study

In authorship related studies, it is often hard to set up an experiment that is entirely 'clean' or 'sterile' from a methodological point of view. If one has to make sure that the *only* difference between two texts is the difference in authorship, one has to keep all other factors (such as gender and education level of the author, topic of the text, ...) as stable as possible over the two texts compared. For the Middle Ages, the poor survival of texts makes it difficult to set up an experiment that fully meets these requirements. For this paper, we shall work on a single case study that does seem to approach this ideal setup as much as possible: the *Spiegel Historiael*, a Middle Dutch adaptation of the Latin *Speculum Maius*, by Vincentius of Beauvais. This adaptation was initiated by the influential writer Jacob of Maerlant and was later continued by two other authors: Filip Uttenbroeke and Lodewijk van Velthem. Of each of these authors near-complete manuscript copies survive of substantial parts of their contribution to the project, called *Partien*. Each of these *partien* is divided in larger units called *books*, which in turn consist of smaller *chapters*. In this study we shall focus on the first *partie* by Maerlant (31K lines in 532 chapters), the second by Uttenbroeke (41K lines in 461 chapters) and the fifth by Velthem (27K lines in 387 chapters). These chapters will be our main comparison unit. What makes this *Spiegel historiael* such an interesting case is that comparing these texts for authorial differences indeed keeps many other factors rather constant, such as level of education, gender, genre, etc.

4. Preprocessing

As a starting point, we shall briefly discuss the architecture which we have developed for the preprocessing of our texts: the MiDL-system (joint work with Walter Daelemans & Guy de Pauw of the Antwerp Computational Linguistics group, CLiPS). The MiDL-system performs tokenization, lemmatization and Part-Of-Speech tagging for Middle Dutch literary texts. The technology we present is optimized for this specific material but should scale well to other medieval languages (or any resource-scarce language characterized by a lot of spelling variation). In this contribution we shall focus on the *corpus-Gysseling* (CG), a corpus that was digitized and semi-automatically annotated at the Institute for Dutch Lexicology (INL). More specifically we shall report on results with the so-called 'literary part' of this corpus (ca. 600K running tokens) that contains all Middle Dutch literature,

surviving from manuscripts dated between 1200 and 1300AD. The main issue we will discuss is lemmatization, as we will argue that this step is actually the key to all subsequent operations (such as e.g. PoS-tagging or shallow parsing).

Lemmatization refers to the process whereby natural language tokens are assigned a 'lemma'. The basic purpose of doing this – in any language or research domain – is that it enables the generalization 'about the behaviour of groups of words in cases where their individual differences are irrelevant' (Knowles & Mohd Don 2004:69). Hence, lemmatization can be considered a problem of mapping many-to-one: similar tokens are mapped to the same 'abstract representation' that, as such, comes to subsume 'all the formal lexical variations which may apply' (Crystal 1997). There exists an obvious parallel with the lexicographer's activity of grouping words under the same 'dictionary headword' (Ibid.; Knowles & Mohd Don 2004:70). When it comes to medieval languages, the main issue that is to be dealt with are historical spelling variants (HSV). When compared to the problem of lemmatization in modern languages, it adds a level of complexity:

Modern

$\text{LEMMA } X = \{\text{token}^1, \text{token}^2, \dots, \text{token}^{n-1}, \text{token}^n\}X$

Medieval

$\text{LEMMA } X = \{\text{token}^1=\{\text{variant}_1^1, \text{variant}_2^1\}, \text{token}^2=\{\text{variant}_1^2, \text{variant}_2^2\} \dots, \text{token}^{n-1}=\{\text{variant}_1^{n-1}, \text{variant}_2^{n-1}\}, \text{token}^n=\{\text{variant}_1^n, \text{variant}_2^n\}\}$

The main purpose of lemmatization, as such, lies with a form of token normalization that allows us to transcend superficial spelling variations.

5. Experiments

In this paper we shall focus on lexical features (n-grams of lemmata) and shallow morpho-syntactic features (n-grams of PoS-tags). Our main research emphasis will be on rhyme words. We will present the results of leave-one-out validation on our data as following. Using a machine learning algorithm, we will do experiments on m samples by Maerlant, n samples by Uttenbroeke and l samples by Velthem (with the sample size set to individual chapter entities). During each fold, we will each time 'leave out' one chapter by one author (e.g. one by Uttenbroeke) and train on the chapters that are left (e.g. m Maerlant-samples, $n-1$ Uttenbroeke-samples and l Velthem-samples). We will 'test' each time the accuracy of our algorithm after training in terms of accuracy: 'Can the learner correctly identify by which author the omitted sample was written?'

References

- Biemans, J.A.A.M.** (1997). *Onsen Speghèle Ystoriale in Vlaemsche. Codicologisch onderzoek naar de overlevering van de Spiegel historiael van Jacob van Maerlant, Philips Uttenbroeke en lodewijk van Velthem, met een beschrijving van de handschriften en Fragmenten*. Leuven: Peeters.
- Crystal, D.** (1997). *A Dictionary of Linguistics and Phonetics*. Oxford: Oxford University Press.
- Dalen-Oskam, K. van and Zundert, J. van** (2007). 'Delta for Middle Dutch – Author and Copyist Distinction in Walewein'. *Literary and Linguistic Computing*. **22**: 345-362.
- Kestemont, M. and Van Dalen-Oskam, K.** (2009). 'Predicting the Past: Memory-Based Copyist and Author Discrimination in Medieval Epics'. *Proceedings of the Twenty-first Benelux Conference on Artificial Intelligence (BNAIC 2009)*. Eindhoven, 2009, pp. 121-128.
- Knowles, G. and Mohd Don, Z.** (2004). 'The notion of a "lemma". Headwords, roots and lexical sets'. *International Journal of Corpus Linguistics*. 69-81.
- Luyckx, K. and Daelemans, W.** (2008). 'Authorship Attribution and Verification with Many Authors and Limited Data'. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*. Manchester, 2008, pp. 513-520.

Building the Humanities Lab: Scholarly Practices in Virtual Research Environments

van den Heuvel, Charles

charles.vandenheuvel@vks.knaw.nl

AlfaLab (Royal Netherlands Academy for Arts and Sciences), Amsterdam, Netherlands; Virtual Knowledge Studio for Humanities and Social Sciences, Amsterdam, Netherlands

Antonijevic, Smiljana

smiljanaantonijevic@vks.knaw.nl

AlfaLab (Royal Netherlands Academy for Arts and Sciences), Amsterdam, Netherlands; Virtual Knowledge Studio for Humanities and Social Sciences, Amsterdam, Netherlands

Blanke, Tobias

tobias.blanke@kcl.ac.uk

King's College London

Bodenhamer, David

intu100@iupui.edu

Polis Center, Indianapolis, USA

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de

University of Wuerzburg, Germany

Nowviskie, Bethany

bpn2f@eservices.virginia.edu

University of Virginia Scholars' Lab, USA

Rockwell, Geoffrey

geoffrey.rockwell@ualberta.ca

University of Alberta, Canada

van Zundert, Joris

joris.van.zundert@huygensinstituut.knaw.nl

AlfaLab (Royal Netherlands Academy for Arts and Sciences), Amsterdam, Netherlands; Huygens Institute, Netherlands

1. Subject

Our Cultural Commonwealth, a report of the American Council of Learned Societies' Commission on Cyber Infrastructure for Humanities and Social Sciences, (ACLS 2006) mentioned the relative "conservative culture of scholarship" in the humanities and social sciences (as compared to the natural sciences) as one of the explanations why researchers in these knowledge domains might be hesitant in using the web and other digital

resources for their research.¹ Before pointing to these researchers however, we may want to question first of all the infrastructures and tools offered to those researchers. Tools and services will only be taken up if they truly *serve* researchers in their daily work. Infrastructures and tools offered to humanities scholars should support concepts and approaches specific to scholarly practices in humanities research, and, to that end, the ACLS commission advocates “working in new ways” by “tools that facilitate collaboration; an infrastructure for authorship that supports remixing, re-contextualization, and commentary—in sum, tools that turn access into *insight and interpretation*. [emphasis in original]” (ACLS, p. 16).

Recently we have seen an increase of virtual laboratories, which use virtual research environments (VREs) to facilitate collaboration among researchers and to promote innovative use of tools and sources in humanities research.²³ Therefore, although traditionally operating as sites of knowledge production in the natural sciences, laboratories have started to develop into loci of scholarly practice in the humanities too. This shift has also been reflected in funding agencies’ support to online laboratory settings; their (re-) allocation of financial resources is fed by expectations that more researchers will become involved in digital and computational humanities, and that the use of information and communication technologies (ICTs) in humanities research will lead to new research questions, methodologies, and ways of collaborating. However, it is not evident that the lab analogy can be transmigrated seamlessly from a science field into the humanities.

Thus far, there has been little critical reflection on such lab initiatives, although the use of VREs in the humanities certainly requires and merits scholars’ attention. Now that a number of such large scale initiatives have developed, there is an opportunity to reflect on what these VREs have achieved and to evaluate their strengths and weaknesses for humanities research, as well as to explore ways in which they should further develop. Therefore this panel addresses theoretical, epistemological, hermeneutical and strategic questions emerging from the use of VREs in digital and computational humanities in general, and in humanities research of text and image in particular. It brings together representatives of scholarly institutions developing virtual labs, infrastructures, and tools to advance the study of text and image in humanities research. The panel focuses on humanities labs promoting new scholarly practices in VREs, and within this broader framework it concentrates on the following specific themes:

- the benefits, challenges and obstacles of research practices emerging from humanities research in virtual research environments.
- the specificities of generating, analyzing and sharing linguistic and visual data in online laboratory settings.
- the advantages and barriers of scholarly collaboration across disciplinary and geographic spans.
- the main features of institutional and funding policies needed for further development of digital humanities labs.
- technical models potentially driving future development of local initiatives.

2. Organization of the panel

The research and development team of AlfaLab⁴ —a digital humanities initiative of the Royal Netherlands Academy of Arts and Sciences (KNAW)— is the initiator and organizer of this panel. Panel members include humanities researchers from KNAW; Digital Research and Scholarship and Scholarly Communication Institute at the University of Virginia Library;⁵ POLIS and Virtual Center on the Spatial Humanities;⁶ TextGRID;⁷ DARIAH;⁸ and the Text Analysis Portal for Research (TAPoR).⁹ The panelists are engaged in the study and development of six different humanities labs dealing with text and image analyses, which grants this panel a unique opportunity to comparatively explore various strategies in building and using humanities labs, and to reflect on both theoretical and practical concerns of that process. In addition, the panel will critically evaluate—from the researchers’ perspective—the focal themes listed above, and it will address actions that might be taken to improve the use of VREs in humanities research.

The panel session will be organized in the following way:

- The panel chair will introduce the main topic, discussion questions, and the panelists; duration: 3 minutes;
- Each of the panelists will give a short presentation (6 minutes), followed by questions from the audience (4 minutes); duration: 60 minutes;
- The themes and questions raised in the presentations will be further discussed in an open forum between the panelists and the audience; duration: 25 minutes;
- The panel chair will briefly reflect on future plans, provide contact information, and close the panel.

Panelists:

- Dr. Tobias Blanke, King's College London discussing DARIAH
- Dr. David Bodenhamer, Director Polis Center discussing the Virtual Center on the Spatial Humanities
- Prof. Dr. Fotis Janidis: University of Wuerzburg (Germany) discussing TextGRID
- Dr. Bethany Nowviskie, Director Digital Research and Scholarship and Associate Director Scholarly Communication Institute at the University of Virginia Library - discussing NINES
- Dr. Geoffrey Rockwell, University of Alberta TAPoR
- Joris van Zundert (MA): Department of Software R&D at the Huygens Institute, Royal Netherlands Academy of Arts and Sciences - project leader AlfaLab, discussing AlfaLab

3. Moderators:

- Dr. Charles van den Heuvel (panel chair)
- Dr. Smiljana Antonijevic Virtual Knowledge Studio (Royal Netherlands Academy of Arts and Sciences)

Notes

1. <http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf>
2. http://mith.umd.edu/tools/?page_id=60
3. <http://www.educause.edu/Resources/UnderstandingInfrastructureDyn/154606>
4. <http://alfalablog.knaw.nl/>
5. <http://nines.org/>
6. <http://www.polis.iupui.edu/>
7. <http://www.textgrid.de/en.html>
8. <http://www.dariah.eu/>
9. <http://portal.tapor.ca/>

Wargames in a Digital Age

Kirschenbaum, Matthew

mgk@umd.edu

English and MITH, University of Maryland

Juola, Patrick

juola@mathcs.duq.edu

Computer Science, Duquesne University

Sabin, Philip

philip.sabin@kcl.ac.uk

King's College London

Wargaming is an applied tradition of interactive modeling and simulation dating back to the early 19th century or, if one counts more abstract martial pastimes like Chess and Go, all the way to antiquity. Why a panel about games (tabletop as well as computer) that spotlight war—surely the most inhumane of organized human endeavor—at a digital humanities conference?

First, we assume that wargaming as both a descriptive or predictive tool as well as a recreational pastime transcends specific technologies of implementation. For example, when a tabletop wargamer moves troops across the battlefield to attack an enemy, they are enacting a specific procedure that is defined against a larger complex of procedures and systems which collectively aspire to represent historical reality within a range of probable (or possible) outcomes. The abstraction of combat, movement, supply, morale, and other basic military considerations into algorithmic process or a numerically expressed spectrum of outcomes—randomized by die rolls within the parameters of a situation—makes the genre a rich source for anyone interested in the formal and procedural representation of dynamic, often ambiguous, literally contested experience.

Second, we are concerned finally not with wargames for their own sake, but as exemplars of simulation as a mode of knowledge representation. As a genre, wargames offer some of the most complex and nuanced simulations in any medium. A typical tabletop game might have many dozens of pages of rules, defining procedures and interactions for hundreds or even thousands of discrete components (unit tokens) across as much as twenty square feet of map space. This places them at the formal and physical extremes of ludic complexity. Almost from the outset of the personal computer revolution, meanwhile, wargames (as distinct from games with superficial militaristic themes) became a major

software genre. Popular tabletop wargames were rapidly translated to the screen by companies such as SSI, with crude artificial intelligence crafting opposing moves. Other games dispensed with the conventions of their manual predecessors and (much like flight simulators) sought to recreate an intense real-time first-person experience. Harpoon (1989) placed a generation of early armchair enthusiasts in the Combat Information Center of a modern naval frigate, with countless variables in weapon and detection systems to master.

We believe that the digital humanities, which have already embraced certain traditions of modeling, might have something to learn from an exploration of this particular genre of simulation, which has proved influential in both professional military and political settings as well as the realm of popular hobby and recreation. (We also find it suggestive that several long-time members of the digital humanities community were “teenage grognards,” suggesting that the games were of a piece with other elements of a particular generational path to computing.)

PAPER 1

Kriegsspiel as Tool for Thought

Matthew Kirschenbaum

mgk@umd.edu

University of Maryland

Kriegsspiel of course is German for (literally) “war game.” In 1824, the Prussian staff officer Georg von Reisswitz formally introduced the game (versions of which had been kicking around in his family for years) to his fellow officers. “This is not a game! This is training for war!” one general is said to have exclaimed (Perla 26). It was quickly adopted, and became the foundation for the German institutionalization of wargaming which persisted through World War II. The von Reisswitz Kriegsspiel was played by laying wooden or metal blocks across maps to mark troop dispositions (Figure 1). Games were conducted on actual topographical maps, often terrain anticipated as the site of future operations (for example, the 1914 Schlieffen plan was subject to extensive rehearsal as a Kriegsspiel). By the middle of the 19th century the “game” had evolved two major variants, so-called “rigid” and “free” Kriegsspiel. The latter attempted to replace the elaborate rules and calculations with a human umpire making decisions about combat, intelligence, and other outcomes on the battlefield.

In this paper, we take the twin traditions of rigid and free Kriegsspiel as our point of departure for thinking about simulation gaming in terms of what Howard Rheingold, in the context of computing, once called “tools for thought.” Indeed, the fork in Kriegsspiel’s development history anticipates much about both manual and computer simulation design. Dungeons and Dragons, the progenitor of all tabletop role-playing systems, grew from a set of medieval wargaming rules called Chainmail. The original developers (Gary Gygax and Dave Arneson) added the magic and monsters, but they also replaced much of the game’s rules apparatus with an umpire dubbed “dungeon master” whose job it was to adjudicate the outcomes of various actions, sometimes with the help of tables and dice, but just as often “freestyle,” relying on judgment and instinct. Wargaming itself has largely remained divided along the same fault between rigid and free systems, with the former attracting hobbyists who buy pre-packaged games (several thousand have been published) to try their hand at Gettysburg or Waterloo and the latter the domain of professional consultants who stage elaborate role playing exercises of the sort originally conducted at thinktanks like RAND but are today as likely to assist a board in planning a corporate merger as a military staff in planning a mission.

As the above suggests, wargames are also both predictive and retrospective in orientation. On the one hand, hobbyist games are often marketed promising insight into the past, tempting a player into believing that with sufficient study and canniness he or she might out-general Napoleon and rewrite history (Dunnigan). In this sense, wargames align with certain strains of academic counterfactual history (Ferguson, et al.). Yet Kriegsspiel was attractive to professional planners precisely because of its predictive value: an accurate formal model of some battlefield dilemma would presumably allow commanders to rehearse their tactics and continually alter the parameters of the situation to arrive at solutions to the military problem. Often, in fact these dual orientations were pursued in tandem, with a historical outcome from a game serving as the control case for subsequent prediction: if a game can restage Midway according to the trajectory of actual events, then in principle outcomes from its hypothetical situations might be equally trusted.

There is yet another way of thinking about wargames though, one that does not assume naïve faith in their capacity as either predictors or descriptors of real-world phenomena. One great virtue of tabletop games is that, by their nature, their rules systems are absolutely transparent. Everything the players need to play the game must be in the box, and the quantitative model underpinning the game system is thereby materially exposed for inspection and

analysis. Many gamers collect and compare dozens of different games on the same subject to see how different designers have chosen to model and interpret events. The hobby is filled with vigorous discussions about designers' intents, as well as house rules and variants, because part of what comes packaged with the game is the game system. (Indeed, the term "game designer" originated at SPI, one of the hobby's premier wargame publishers.) As one wargame enthusiast shrewdly observes, "What wins a wargame is but a dim reflection of what wins a battle, or a war. Sometimes, what wins a wargame doesn't reflect reality at all" (Thompson). In this view, the game engine is a procedural instrument for producing an outcome whose value lies in its potential for provoking counter-factual analysis. A wargame--either manual or computer--may permit Napoleon to win at Waterloo: the salient question is not whether the game was "right" but in the questions it exposes about whether Napoleon really could have done so (and if so, how). This viewpoint actually comports with that of professional wargame facilitators, who assert that the ultimate value of their games is *not* predictive in any simple sense, but rather as "part of a process persuading people that there are other ways to think about problems" (Herman 59). A modern boardroom wargame, in other words, provides a safe space in which participants can explore solutions that would not have been ventured in a more conventional setting.

After establishing this background through examples, the paper will propose a new Kriegsspiel implementation that modulates between rigid and free design parameters in order to expose—deliberately—the workings of the game engine as a tool for the kind of thinking Thompson suggests. The key counter-factual analysis is access to the game's internal systems, and analysis of their function as systems for procedural representation, what Kirschenbaum has elsewhere called *procedural granularity*. Our Kriegsspiel model will thus permit play of the game in its various historical incarnations, while simultaneously exposing and even directing user attention to various game systems. At the same time, our model draws inspiration from von Reisswitz's attempt to simulate the "fog of war." This term, which was coined by that most influential of all modern military theorists, Carl von Clausewitz, aptly describes the gaps in situational awareness experienced by soldiers and commanders on the battlefield. We note, however, that it also corresponds to the more modern game theoretic notion of "imperfect information" and to the general idea that successful simulation—both as analytical exercise and as imaginative activity—depends largely on what is not "filled in" by the game environment. We believe, indeed, that by building

these kinds of environments, we can come to a better understanding of how this important dynamic works in interactive environments more generally. The kind of Kriegsspiel we propose is finally a tool not for thinking about war, but for thinking about representation and design.



Figure 1. A game of Kriegsspiel played using a modern set

References

- Clauswitz, Carl von** (2008). *On War*. Oxford: Oxford University Press.
- Dunnigan, James F.** (1992). *The Complete Wargames Handbook*. New York: Quill.
- Ferguson, Niall (ed.)** (1999). *Virtual History: Alternatives and Counterfactuals*. New York: Basic Books.
- Herman, Mark, Frost, Mark, Kurz, Robert** (2009). *Wargaming for Leaders*. New York: McGraw Hill.
- Kirschenbaum, Matthew G.** (2009). "War Stories: Board Wargames and (Vast) Procedural Narratives". *Third Person*. Harrigan, Pat, Wardrip-Fruin Noah (eds.). Cambridge: MIT Press, pp. 357-71.
- Leeson, Bill (trans.)** (2007). *The von Reisswitz Kriegsspiel*. Too Fat Lardies.
- Perla, Peter** (1990). *The Art of Wargaming*. Annapolis: Naval Institute Press.
- Thompson, Nels** (2009). 'Learning from Wargames'. *Battles*. 2 : 60.

PAPER 2

What Does It Feel Like When They Put you Back in the Box? : Representation and Mathematics in Tactical Simulations

Patrick Juola

juola@mathcs.duq.edu
Duquesne University

Simulation is a key method for analyzing situations and events as well as for presenting them to the public. Although this panel focuses on the simulation of military events (Arnhem), the same principles apply to computer-based simulations as well as personal “role-playing” games, and to simulations of other types, such as financial simulations (1830), medical simulations (ER; see also Halloran et al, 2009), or political situations (Origins of World War II).

In broad terms (Frasca 2001; Kirschenbaum 2003), a simulation is a narrative generator, a system containing in potential a large number of possible sequences of events. At the same time, to be practical, a simulation must quantize the infinite variety of potential narrative reality to a small set of event categories, a set small enough to be tractable and manipulable to the players. A simple example of this is the playing field or map itself. In a typical tactical simulation, the map will be “discretized” into a regular array of hexagonal regions, typically assumed to uniform in composition (“forest hexes”), possibly with edge effects such as rivers or walls. Another example is the with edge effects such as rivers or walls. Another example is the playing piece itself, which can range from a simple wooden counter (as in Risk or Diplomacy) to a bewildering array of symbols representing high level abstract properties of a multi-person combat unit (Arnhem), or even a detailed schematic of individual functional capacities (Star Fleet Battles).

Similarly, the relationship of events to each other is controlled by game rules describing the set of permissible actions and their (possibly probabilistic) outcomes. For example, “ships” are not typically permitted to move through “forest”; “cavalry” usually moves faster than “artillery,” and the effect of “encountering” enemy artillery may result in the elimination of a counter, its enforced movement (“retreat”), or other effects.

In this paper, we analyze the mathematical basis for these representations, stripping them both of their narrative aspects (the association of any particular hex with the Argonne forest, for example) as well as their technological aspects (whether the region is represented by colored cardboard, pixels, or an abstract name). We focus particularly on the differences between quantified and non-quantified representations as well as between probabilistic and deterministic representations. We also discuss some of the aspects of the unrepresented—and therefore illegal—aspects of reality. In some cases, these can be seen as aspects of increasing realism by disallowing activities that could not physically take place in our hypothetical universe, but can also be seen as limiting the choices for a creative player, or even of enforcing some sort of political correctness upon the game universe itself by outlawing possible but distasteful alternatives. We suggest both that the narratives generated as well as our analysis of simulated narratives can be enhanced by an understanding of the abstract structure of the representations, and that this may eventually enhance our ability to understand non-simulated narratives such as those generated by counterfactual historians.

References

- Ferguson, Niall (ed.)** (1999). *Virtual History: Alternatives and Counterfactuals*. New York: Basic Books.
- Frasca, Gonzalo** (2001). "SIMULATION 101: Simulation versus Representation". <http://www.ludology.org/articles/sim1/simulation101.html>.
- Halloran M.E., Ferguson N.M., Eubank S., Longini I.M., Cummings D.A., Lewis B., Xu S., Fraser C., Vullikanti A., Germann T.C., Wagener D., Beckman R., Kadau K., Barrett C., Macken C.A., Burke D.S., Cooley P.** (2008). 'Modeling targeted layered containment of an influenza pandemic in the United States'. *PNAS*. **105:1073**.
- Kirschenbaum, Matthew** (2003). 'I was a Teenage Groggnard'. <http://atal.umd.edu/~mgk/blog/archives/000235.html>.
- Salen, Katie, Zimmerman, Eric** (2004). *Rules of Play : Game Design Fundamentals*. Cambridge: MIT Press.
- (1983). *1830*. Avalon Hill.
- (1972). *Arnhem*. Panzerfaust Publications.
- (1959). *Diplomacy*. Avalon Hill.
- (2005). *ER*. Vivendi Interactive.

- (1971). *Origins of World War II*. Avalon Hill.
- (1957). *Risk*. Parker Brothers.
- (1979). *Star Fleet Battles*. Task Force Games.

PAPER 3

The Benefits and Limits of Computerisation in Conflict Simulation

Philip Sabin

philip.sabin@kcl.ac.uk
King's College London

Ever since the development of 'Kriegsspiel' nearly two centuries ago, military professionals and enthusiasts have used simulation and gaming techniques to model real military conflicts.¹ This phenomenon builds on the theoretical similarity between war and games, in that both are dialectical strategic contests between opposing wills, each struggling to prevail.² Hence, Clausewitz said that 'In the whole range of human activities, war most closely resembles a game of cards'.³

The growing potential of computers has naturally transformed the field of conflict simulation. Military training now employs networked computer arrays running real time first person models of entire conflict environments, and millions of enthusiasts use similar first person simulations of air combat, ground fighting and the like.⁴ However, what is interesting is the persistence of traditional manual simulation techniques alongside this computerised mainstream. Just as military forces continue to use real field exercises, so many enthusiasts continue to employ pre-computer age techniques such as maps and counters in their modelling of conflict. Indeed, such 'manual' wargames are now being published at a faster rate than ever before, and there are still far more manual than computer simulations in existence, especially of historical conflicts.⁵

I have been playing and designing conflict simulations for over three decades, and I use both manual and computerised versions routinely as instructional aids and research tools in the War Studies Department at KCL, including through an MA course in which students design their own simulations of conflicts of their choice.⁶ In this paper, I will explore the benefits and limits of computerisation in conflict simulation, and explain why my own forthcoming book *Simulating War*

focuses so heavily on manual simulation techniques despite the ongoing computer revolution.

The paper will have two central themes. One is the complex and double-edged nature of 'accessibility' in the simulation field. Computer simulations tend to be more accessible to users, but harder to programme and design, so they are best suited to expert-led situations in which a few highly capable individuals devote considerable effort to creating a model which can be learnt and used 'as is' by masses of less qualified people. Manual simulations, by contrast, tend to be less accessible to users because they need to master lengthy rules to be able to operate the model at all, but in the process the users are required to engage much more directly with the designer's ideas and assumptions, and it is a short step from being able to play a manual simulation to being able to tweak the rules or even to design entirely new systems to give a better reflection of one's own understanding of the underlying military reality. Hence, manual simulations are much more accessible from a design perspective, since one does not need to be a computer programmer to create new systems, and since using other people's systems conveys a much better understanding of design techniques.

Many recent computer simulations have sought to soften their expert-led character by incorporating provision for simple modification and scenario generation by users themselves.⁷ However, this flexibility rarely extends to changing the fundamental systems, and it is actually manual simulation design which has become radically more accessible and democratised in the computer age, thanks to the ease with which individuals can now design full colour maps and counters and sell or give away digitised copies of their rules and graphics online without any physical production or distribution costs.⁸ Since I believe that designing simulations for oneself is a far better way of gaining insight into the dynamics of a real conflict than is simply playing someone else's computer game on that subject, I see the much greater design accessibility of manual simulations as a major reason for their continued production and relevance, with computer graphics and online distribution playing a key role, but without the rules themselves having to be coded into computer software.

The second key theme of this paper will be that the relative advantages of manual and computer simulation vary greatly depending on the type of conflict being modelled and the perspective which users are intended to adopt. Broadly speaking, the more fast-moving and physically calculable the conflict environment, and the more that users are intended to experience the perspective of a single real individual, the more that computers have to

offer. Hence, although it is possible to simulate aerial dogfights using maps, counters and dozens of pages of highly complex and time-consuming rules, the fast-paced 3D manoeuvres are obviously much better captured by real-time computer simulations from the perspective of the individual cockpits, and this is exactly what I use in my own teaching about air combat.⁹ Even when simulations are intended to model entire battles, computers can employ AI routines to mimic the limited perspectives of an individual commander, by masking the full picture in a way which manual simulations find harder because their users must run the whole system rather than just playing individual roles within it.¹⁰

The trouble with computers is that their unparalleled number-crunching abilities tend to encourage the dangerous belief that accurate simulation is primarily a matter of adding more and more parameters and increasingly detailed data. Manual simulation designers, by contrast, must perform focus on identifying and modelling the really significant dynamics in that particular conflict, since their games would otherwise be completely unplayable.¹¹ This pushes them more towards an output-based, top-down design approach, whereas computer programmers tend to prefer more input-based, bottom-up techniques. The differences can be striking. For instance, networked first person computer simulations of infantry combat tend to produce grossly ahistorical casualty rates despite highly precise and detailed modelling of terrain and weaponry, because the individual participants behave far more boldly than they would if the bullets were real. Manual simulations find it much easier to model this suppressive effect of fire, by simply prohibiting users from moving troops who are pinned down in this way.¹² Since it is very common indeed for conflicts to be affected at least as much by such psychological dynamics as by more calculable physical parameters, manual simulations can often identify and capture the 'big picture' at least as effectively as do apparently more detailed computer models.¹³

The central message of this paper will be that 'simulation' and even 'digitisation' are not necessarily synonymous with 'computerisation', as so many today seem to believe. Military professionals and enthusiasts have been producing 'digitised' mathematical models of conflict since long before the computer age, and such manual simulations continue to flourish alongside their computerised counterparts. The biggest challenge they face is that computer simulations now have much greater mass market appeal and a much more professional image within defence and academia. However, without the broad accessibility and top-down focus of manual simulation design, computerised conflict simulation

would become an unduly arcane and detail-obsessed science. Manual and computer simulations of conflict will hence remain complementary endeavours for many years to come.

Notes

1. See P.Perla, *The Art of Wargaming*, (Annapolis: Naval Institute Press, 1990) and J.Dunnigan, *The Complete Wargames Handbook* (New York: William Morrow, 2nd ed., 1992).
2. See T.Cornell & T.Allen (eds.), *War and Games*, (Rochester NY: Boydell, 2002), which includes a chapter by myself.
3. C. von Clausewitz, *On War*, edited and translated by M.Howard & P.Paret, (Princeton: Princeton University Press, 1976), p.86.
4. The similarity between the genres has become so great that virtually the same games are often employed by military professionals and civilian enthusiasts, as with the commercial game *Armed Assault* (Bohemia Interactive, 2007), whose military variant *VBS2* is widely used as a training aid and has now even been released back to the public by the UK Ministry of Defence as a recruitment device!
5. See the flood of new manual game announcements on <http://www.consimworld.com>, and compare this with the survey 30 years ago in N.Palmer, *The Comprehensive Guide to Board Wargaming*, (London: Arthur Barker, 1977) and with the new computer game announcements on <http://www.wargamer.com>.
6. See my course website at <http://www.kcl.ac.uk/schools/sspp/ws/people/academic/professors/sabin/conflictsimulation.html>, and my book *Lost Battles: Reconstructing the Great Clashes of the Ancient World*, (London: Hambledon Continuum, 2007).
7. See, for example, *Armed Assault* (Bohemia Interactive, 2007), and Norm Koger, *The Operational Art of War III*, (Matrix Games, 2006).
8. See, for instance, <http://www.wargamedownloads.com> and <http://cyberboard.brainiac.com/>.
9. Compare, for example, J.D.Webster's manual game *Achtung – Spitfire!*, (Phoenixville PA: Clash of Arms, 1995), with the PC game *Battle of Britain II: Wings of Victory*, (G2 Games, 2005).
10. See, for instance, the PC games *Take Command: Second Manassas*, (Paradox Interactive, 2006), and *Airborne Assault: Conquest of the Aegean*, (Panther Games, 2006).
11. The classic example of such an unplayable monster is Richard Berg's *Campaign for North Africa*, (New York: Simulations Publications Incorporated, 1979).
12. This is well illustrated in Phil Barker, *War Games Rules, 1925-1950*, (Wargames Research Group, 1988).
13. See my book *Lost Battles*, (London: Hambledon Continuum, 2007), and the manual simulation which I co-authored with my former MA student Garrett Mills on *Roma Invicta? Hannibal in Italy, 218-216 BC*, (Society of Ancients, 2008). I use both of these in my teaching on ancient warfare, and I use similar manual simulations in my classes on the operational and strategic aspects of modern warfare.

Scanning Between the Lines: The Search for the Semantic Story

Lawrence, K. Faith

f.lawrence@ria.ie

Royal Irish Academy

Battino, Paolo

p.battino@dho.ie

Royal Irish Academy

Rissen, Paul

paul.rissen@bbc.co.uk

British Broadcasting Corporation

Jewell, Michael O.

m.jewell@gold.ac.uk

Goldsmiths College, University of London

Lancioni, Tarcisio

lancioni@unisi.it

University of Siena

The panel will present three projects which are exploring the use of metadata to describe the narrative content of media. Computer-assisted textual analysis is now a well known and important facet of scholarly investigation (Potter, 1991; Burrows, 2004; Yang, 2005) however it relies heavily on statistical approaches in which the computer uses character matching to identify reoccurring strings. Although pattern recognition for image and audio search is growing more sophisticated (Downie, 2009), the techniques for annotation of multimedia are subject to the same limitations as those for text in that they cannot go beyond the shape or the waveform into the meaning that those artifacts of expression represent.

This limitation has been addressed in a number of different ways, for example through traditional categorisation and with the use of keywords for theme and motif annotation. New techniques using natural language processing software such as GATE (Auvil, 2007), and IBM's LanguageWare (1641 Depositions Project - <http://www.tcd.ie/history/1641/>) have taken this further allowing a deeper level of meaning to be inferred from the text through basic entity and relationship recognition. The use of ontologies to support this annotation opens the way for more precise search, retrieval and analysis using the techniques developed in conjunction with semantic and linked web architecture.

The three papers being presented in this panel will address the application of these techniques to both textual and audio-visual media and consider annotation not just of the documents themselves but of the ideas contained within them, how this information might be presented to the user to best effect.

The first paper in this panel by Dr Michael O. Jewell, Goldsmiths College, University of London, focuses on the annotation of narrative in scripts and screenplays. This paper presents the combination of TEI and RDF annotations as a methodology for opening the encoded data up for inference-enhanced exploration and augmentation through linked-data resources.

The second paper from Paul Rissen, BBC, and Dr K Faith Lawrence, Royal Irish Academy, presents work being done at the BBC in the annotation of the narratives within their audio/visual archives. This paper discusses the initiatives within the BBC to make their content more accessible and to allow more personal interaction with the material. Through the use of ontology, the events contained the media object are exposed to exploration, analysis and visualisation.

The final paper by Paolo Battino, Royal Irish Academy, continues the visualisation theme to discuss how narrative annotations might be presented to assist in analysis of texts. Using the example of folktales, this paper considers the graphical representation of plotlines and the possible issues and challenges inherent for visualisation in moving from syntactic to semantic description.

References

Auvil, L., Grois, E., Lloràname, X., Pape, G., Goren, V., Sanders, B., Acs, B., McGrath, R. E. (2007). 'A Flexible System for Text Analysis with Semantic Networks'. *Proceedings of Digital Humanities 2007*.

Burrows, J. (2004). 'Textual Analysis'. *A Companion to Digital Humanities*. Schriebman, S., Siemens, R., Unsworth, J. (eds.). Oxford: Blackwell Publishing Ltd.

Downie, J. S., Byrd, D., Crawford, T. (2009). 'Ten Years Of Ismir: Reflections On Challenges And Opportunities'. *10th International Society for Music Information Retrieval Conference*.

Potter, R. G. (1991). 'Statistical Analysis of Literature: A Retrospective on Computers and the Humanities, 1966–1990'. *Computers and the Humanities*. **25.6**: 401-429.

Yang, H-C., Lee, H-C. (2005). 'Automatic Category Theme Identification and Hierarchy Generation for Chinese Text Categorization'. *Journal of Intelligent Information Systems*. 1.

PAPER 1

Semantic Screenplays: Preparing TEI for Linked Data

Jewell, Michael O.

m.jewell@gold.ac.uk

Goldsmiths College, University of London

Scripts, whether for radio plays, theatre, or film, are a rich source of data. As well as cast information and dialogue, they may include performance directions, locations, camera motions, sound effects, captions, or entrances and exits. The TEI Performance Texts module (<http://www.tei-c.org/P5>) provides a means to encode this information into an existing screenplay, together with more specific textual information such as metrical details.

Meanwhile, Linked Data has become a major component of the Semantic Web. This is a set of best practices for publishing and connecting structured data on the Web, which has led to the creation of a global data space containing billions of assertions, known as the Web of Data (Bizer et al, 2009). Some of the most prominent datasets in this space include DBpedia, with more than 100 million assertions relating to (amongst others) people, places, and films; LMDB (Linked Movie Database), with over three million filmic assertions; and LinkedGeoData, which has almost two billion geographical assertions.

In this paper, we propose a means to support Linked Data in TEI, thus benefitting from the wealth of information available on top of that which is provided by TEI. We describe the augmentation of TEI documents with RDFa (Resource Description Format in Attributes) to complement the annotated content with URIs and class information, and thence the transformation of this document into triples using our open source teizonto conversion tool. Finally, we provide some case studies that make use of the resultant triples, and show how their compliance with the OntoMedia ontologies (Lawrence et al, 2006) allows for powerful research possibilities.

1. Annotating TEI

1.1. Cast Lists

The first, and simplest, step to adding RDFa attributes to a TEI document begins with the cast list. Listing 1 shows a simple example of a castItem element for the role of Jeffrey Beaumont in Blue Velvet, portrayed by Kyle MacLachlan. The about attribute specifies the object to which the element relates: the actor element refers to the DBpedia entry for Kyle MacLachlan, while the role refers to an object residing within the Blue Velvet namespace, created specifically for this screenplay. The **property** attribute defines the predicate that relates the content of the element to the object - in this case, it is the name of the actor or character. When processed by **teizonto**, actors are specified as **Being** objects, which are subclasses of the Friend of a Friend (FOAF) ontology's Agent class (<http://www.foaf-project.org/>), and roles as Character objects.

The conversion script then analyses **sp** elements for who attributes. These refer to the **xml:id** attributes in the role elements, and thus it is possible to determine the cast present in a scene and the entity speaking a line. The former may be found via the involves predicates of the event, while the latter is represented with the **has-subject-entity** predicate. An OntoMedia Social event is created for each element, with the **precedes** and **follows** attributes describing the sequence of these events in the screenplay. Listing 2 shows the N3 representation created from a single sp element. Listing 2: The N3 extracted from a TEI sp element given an annotated castList and valid who attributes. The ome prefix refers to the OntoMedia Expression namespace.

```
<castItem>
  <actor
    about="[dbpedia:Kyle_MacLachlan]"
    property="foaf:name">Kyle MacLachlan</
  actor>
  <role
    xml:id="jeffrey_beaumont"
    about="[bv:Jeffrey_Beaumont]"
    property="foaf:name">Jeffrey Beaumont</
  role>
</castItem>
```

Listing 1: A TEI **castItem** augmented with RDFa.
The **dbpedia**, **foaf** and **bv** prefixes refer to the DBpedia, FOAF, and custom Blue Velvet namespaces.

```
<http://contextus.net/resource/blue\_velvet/
event/6>
a ome:Social; ome:follows
<http://contextus.net/resource/blue\_velvet/
event/5>;
```

```

ome:has-subject-entity bv:Detective_Williams;
ome:involves
bv:Detective_Williams, bv:Jeffrey_Beaumont;
ome:precedes
<http://contextus.net/resource/blue_velvet/
event/7>;

```

Listing 2: The N3 extracted from a TEI **sp** element given an annotated **castList** and valid who attributes. The **ome** prefix refers to the OntoMedia Expression namespace.

1.2. Locations

OntoMedia provides an extensive location ontology, and it is thus useful to be able to specify this within a TEI document. **teizonto** analyses the stage elements for this purpose, specifically when the type attribute is given as 'location'. The only compulsory attribute is about which, as with the castList elements, typically refers to an object in the screenplay's namespace. This allows for references to the same location several times throughout a screenplay, or even for other screenplays to reference it (e.g. the Statue of Liberty, or Area 51).

By default, **teizonto** defines locations as being instances of the Space class. This is the highest level Location class, and equivalent to the AKT Location Ontology Abstract-Space class. To specify a more relevant class, the RDFa **typeof** attribute may be used. Furthermore, by specifying a p sub-element it is possible to use the textual name of the location as the dc:title of the object. Listing 3 gives an example of this, with Listing 4 showing the generated location N3 representation for the event from Listing 2 which is set in Room 221.

2. Expression References

Finally, we augment the **rs** element with the RDFa about attribute to provide a powerful means of object reference. OntoMedia defines the **refers-to** property as a means to indicate that an Expression object refers to another Expression object. For example, an event may refer to a location, or a character, or even another event. By adding this attribute to a TEI document many interesting queries may be performed. Listing 5 illustrates a (slightly abbreviated) set of examples of this, and Listing 6 contains the full N3 result generated for the two **sp** elements.

```

<stage type="location" about="[bv:Room_221]"
typeof="[loc:Room]">
  <p property="dc:title">INT. ROOM 221 -
  POLICE STATION</p>
</stage>

```

Listing 3: A TEI stage direction augmented with RDFa

```

bv:Room_221 a loc:Room;

```

```

dc:title "INT. ROOM 221 - POLICE
STATION"^^<rdf:XMLLiteral>.

<http://contextus.net/resource/blue_velvet/
event/6> a
ome:Social;
loc:is-located-in bv:Room_221.

```

Listing 4: The N3 generated from Listing 3

```

<sp who="#jeffrey_beaumont">
  <speaker>JEFFREY</speaker>
  <l>Is <rs type="person"
about="[bv:Detective_Williams]">Detective
Williams</rs> here?</l>
</sp>
<sp who="#desk_sergeant">
  <speaker>DESK SERGEANT</speaker>
  <l>He's up in <rs type="place"
about="[bv:Room_221]">Room 221</rs>. </l>
</sp>

```

Listing 5: Augmenting the rs element to provide location and character references

```

<http://contextus.net/resource/blue_velvet/
event/1> a ome:Social;
  ome:has-subject-entity bv:Jeffrey_Beaumont;
  ome:involves bv:Desk_Sergeant,
  bv:Jeffrey_Beaumont;
  ome:precedes
<http://contextus.net/resource/blue_velvet/
event/2>;
  ome:refers-to bv:Detective_Williams;
  loc:is-located-in
  bv:Police_Department_Reception.

```

```

<http://contextus.net/resource/blue_velvet/
event/2> a ome:Social;
  ome:follows <http://contextus.net/resource/
blue_velvet/event/1>;
  ome:has-subject-entity bv:Desk_Sergeant;
  ome:involves bv:Desk_Sergeant,
  bv:Jeffrey_Beaumont;
  ome:precedes
<http://contextus.net/resource/blue_velvet/
event/3>;
  ome:refers-to bv:Room_221;
  loc:is-located-in
  bv:Police_Department_Reception.

```

Listing 6: The N3 generated from Listing 5

```

PREFIX bv: <http://contextus.net/resource/
blue_velvet/>
PREFIX ome: <http://purl.org/ontomedia/core/
expression#>

```

```

SELECT DISTINCT ?event WHERE
{
  ?event ome:refers-to bv:Detective_Williams.
}

```

Listing 7: SPARQL to retrieve every scene referring to Detective Williams.

As some of the objects in the screenplay refer to external entities, it is also possible to make use of Linked Data. For example, DBpedia has a great deal

of information regarding Kyle MacLachlan, who we have specified as the actor playing Jeffrey Beaumont. Listing 8 gives an example of one of these more powerful queries. In this query, every actor starring in Blue Velvet is retrieved and then every other film that they have starred in is retrieved from DBpedia. The directors of these films are then obtained as URIs. Other queries could, for example, find the most common nationality among cast members, or find actors who have been in other films by the same director. Furthermore, the OntoMedia structure could be leveraged to find films in which the same actors have played alongside each other. For series, the character URIs could also be incorporated - for example, to find every episode in which a character has had a scene set in a factory.

3. Conclusion

The **teizonto** TEI translation tool provides a quick and non-intrusive approach to make use of the Web of Data's Linked Data. Even with the simple addition of about attributes on the cast list, hundreds of assertions about the actors are immediately available. Once location references are provided, it is possible to analyse location usage through TV series, or link them to their real-life counterparts to examine setting information. Finally, the rs element allows for interlinking within events: be it to find where characters and locations are introduced, to investigate which characters are the most talked about, or even to find references to characters in an entirely different film.

```
PREFIX bv: <http://contextus.net/resource/blue_velvet/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX omb: <http://purl.org/ontomedia/ext/common/being#>

SELECT DISTINCT ?director WHERE
{
  ?character omb:portrayed-by ?actor .
  ?film dbpedia:starring ?actor ;
  dbpedia:director ?director .
}
```

Listing 8: SPARQL to find people who have directed the Blue Velvet actors in other films.

Future revisions of **teizonto** will include support for TEI person, place, and trait data; an approach to represent references to the original TEI document via XPath; and more specific camera and movement description. Finally, we will be providing annotated TEI and the accompanying N3 and RDF at the Contextus Data Store website (<http://contextus.net/~datastores>), which we hope will provide an entry point into the Semantic Web for narrative researchers.

References

- Bizer, C., Heath, T., Berners-Lee, T.** (2009). 'Linked Data - The Story So Far'. *International Journal on Semantic Web and Information Systems (IJSWIS)*.
- Harris, S., Lamb, N., Shadbolt, S.** (2009). '4store: The Design and Implementation of a Clustered RDF Store'. *The 5th International Workshop on Scalable Semantic Web Knowledge Base Systems*.
- Lawrence, K. F., Jewell, M. O., Schraefel, M. C., Prugel-Bennett, A.** (2006). 'Annotation of Heterogenous Media Using OntoMedia'. *First International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*.

PAPER 2

Re-imagining the Creative and Cultural Output of the BBC with the Semantic Web

Rissen, Paul

paul.rissen@bbc.co.uk
British Broadcasting Corporation

Lawrence, K. Faith

f.lawrence@dho.ie
Royal Irish Academy

This paper will introduce the work being done at and in conjunction with the BBC into using descriptive metadata to improve production and distribution of content but in such a way that it is positioned within the greater cultural context. With the increased digitisation and release of resources there is also an increased need for associated information that can be computer processed and analysed to adequately index and search the rapidly expanding pool of data. The use of standard metadata formats for description and storage is now part of good practice but is still limited in its scope and application. In this paper we will discuss the ongoing research into semantic description of media content to supplement and expand on current metadata practice to allow more detailed analysis and visualisation of digitised documents and the conceptual links between them.

1. New Media, New Opportunities

For the past twenty years, the World Wide Web has been used by media companies as an enabling technology, allowing them to do the same things as they have always done, but faster and cheaper, whilst making their content more widely available, and for longer periods of time. In addition, the Web has been used as a promotional and marketing tool, increasing the public awareness of content, and providing direct consumption opportunities. In the UK, this can most obviously be seen through the successes of the BBC's iPlayer (<http://www.bbc.co.uk/iplayer/>), and Channel 4's 4 On Demand (<http://www.channel4.com/~programmes/4od>) services. However, despite these successes, it can be argued that the tendency is for media companies not to have taken full advantage of the creative opportunities offered by web technologies due to concerns regarding control of content, licensing issues, technical limitations and the need to ensure a revenue stream.

There has been much discussion and research within and around the industry on the topic of interactivity analysing of how the web could be used to offer new, more immersive and satisfying experiences to the audiences. The EU's NM2 'New Millennium, New Media' project (<http://www.ist-nm2.org/publications/deliverables/deliverables.html>) commissioned a number of studies on audience appreciation of media output in addition to paving the way for various experiments into the use of traditional IT-based production tools to enhance existing content offerings (Ursu, 2008). However, we argue that this work has been heavily based in the traditional understanding of media production and as such uses graphical technologies, such as Flash, whereby the audience is still reduced to a primarily passive role in the consumption of media content. This leads to a mis-use of the term 'interactive' to indicate not a more personalised experience but one which is being delivered through a different medium.

2. Interactivity and the Semantic Document

The hype surrounding the term 'Semantic Web' (Berners-Lee et al., 2001) in recent years has led some to doubt its existence and the opportunities it claims to offer. Although commonly acknowledged to be a complex subject, at its heart, the idea of the semantic web, and of the web itself, is simple. Concepts which are of interest, be they people, places, events or cultural movements, are given unique, permanent identifiers, and links, crucially links with meaning, are drawn between them. In this, its roots can be seen very much in the original design of the Web as put forward by Sir Tim Berners-Lee.

The ideas inherent in the Semantic Web reflect the experience of our own understanding of the world, where, it could be argued, it is the context, i.e. the links we make in our minds, rather than individual objects themselves, which are of the most value. The Semantic Web seeks to replicate this in digital form, and improve upon it, by making these links solid, permanent, and recorded.

At the BBC, three recent projects - BBC/programmes, BBC/music and BBC/wildlifefinder - have sought to apply these principles to parts of our output. These projects sought to increase the value for the content for the user-audience by allowing them to create their own journeys across the resources made available to them on the BBC website and at the same time drawing increased understanding and insight from the knowledge presented on those pages and from selected external sources across the web.

The work to date has concentrated on the structures of production and distribution of BBC content. However, we argue that the real audience value and appreciation can be gained by applying the same approach to the content itself and annotating not only the video/audio files we create, but, more importantly, the narrative structures contained within them. Research into audience and fan studies (for instance Jenkins, 1992; Harris et al., 1998; Baym, 2000; Hills, 2002; Jenkins, 2006) suggests that the audience is creating and expanding a narrative structure within their minds while they are watching media content. While, in reference to this research, this factor was seen as applicable to the content produced by the BBC, its wider relevance should be noted.

Building on this initial work, our research uses a RDF-based triplestore and the OntoMedia ontology to recreate, using semantic web technologies, the users experience of narrative within media. The OntoMedia ontology was designed at the University of Southampton to enable expression of narrative structures within and between mono- and multimedia documents (Lawrence, 2007, Part IV).

3. The Semantic Viewer

In the context of the BBC's drama output, or by corollary any similar corpus of work, we chose to apply the methodology described above as part of a pilot project. By giving each character, location and significant plot event an identifier, and creating meaningful links between them, in parallel with those found in the media itself, we argue that we can allow audiences to follow their own path through the narrative, to explore stories from different points of view, and to achieve a greater appreciation, and true interaction with, the writers' craft. In the diagram shown below (Fig. 1), the events of the Doctor

Who episode Blink (Steven Moffat, 2007) are linked not only the narrative order in which they were experienced within the broadcast but also within the chronological order within the fictional universe containing them and to the orders in which specific characters perceived them. This information, once described with the OntoMedia ontology, is stored within the repository as triples where it can be queried, analysed and visualised.



Fig. 1: Timelines For Doctor Who Episode Blink

The potential of this approach can also be seen in greater terms when applied to other areas of a media company such as the BBC's traditional output. Documentaries are often forms of narrative that seek to draw meaningful links between diverse concepts, in order to educate and inform audiences. These same audiences find enjoyment in our coverage of sport precisely because they are able to place what we report in a wider, linked context. Even news coverage, when seen through this lens, could be transformed, allowing audiences to construct a better understanding of the world we live in, by seeing things from multiple points of view, and constructing their own, informed opinions on events.

4. Conclusion

While the research we discuss in this paper deals with the annotation of narrative within audio/visual media the techniques discussed have a much wider ranging applications. The ontology in use was designed to deal with multiple types of sources and in conjunction with other metadata standards such as CIDOC CRM (<http://cidoc.ics.forth.gr/>), FRBR (<http://www.frbr.org/>) and TEI (<http://www.tei-c.org/>). This allows for the conceptual links between many different types of documents to be made explicit in such a way that a computer could analyse, process and visualise them. While a consumer of BBC content might wish to interrogate the narrative from different perspectives, so might a literary scholar or a historian wish to explore their sources were these techniques applied to heritage materials.

References

- Baym, N.** (2000). *Tune In, Log On: Soaps, Fandom and Online Community*. California: Sage Publications.
- Berners-Lee, T., Hendler, J., Lassila, O.** (2001). 'The Semantic Web'. *Scientific American Magazine*.
- Harris, C., Alexander, A.** (1998). *Theorizing Fandom: Fans, Subculture and Identity*. New Jersey: Hampton Press, Inc.
- Hills, M.** (2002). *Fan Cultures*. London: Routledge.
- Jenkins, H.** (1992). *Textual Poachers: Television Fans and Participatory Culture*. London: Routledge.
- Jenkins, H.** (2006). *Fans, Bloggers, and Gamers: Exploring Participatory Culture*. New York: New York University Press.
- Lawrence, K. F.** (2007). *The Web of Community Trust - Amateur Fiction Online: A Case Study in Community Focused Design for the Semantic Web*, Doctoral Thesis, University of Southampton.
- Ursu, M. F., Thomas, M., Kegel, I., Williams, D., Lindstedt, I., Wright, T., Leurdijk, A., Zsombori, V., Sussner, J., Myrestam, U., Hall, N.** (2008). 'Interactive TV Narratives: Opportunities, Progress, and Challenges'. *ACM Trans. Multimedia Comput. Commun. Appl.*. Tuomola, M. (ed.). **4**, 4, Oct.

PAPER 3 Visualization and Narrativity: A Generative Semiotics Approach

Battino, Paolo

p.battino@dho.ie
Royal Irish Academy

Lancioni, Tarcisio

lancioni@unisi.it
University of Siena

Storing text in a digital format opened up a whole range of new possibilities of unconventional visualization techniques, including displaying texts in forms other than textual. Similarly to what happens with pie charts and histograms in representing large amounts of numeric data, word

clouds and visual taxonomies allow grasping at a glance some interesting relationships among text's constituents. The power of these graphic representations of e-texts is often based on the possibility of counting occurrences of each word, clustering words, identifying recurring patterns of words.

However, as we move from syntax to semantic, we may run into some serious limitations. Comparatively, there is a limited number of text visualisation tools based on semantics of words. Few tools, if any, exist to account for deeper semantic structures underlying texts. Finding a tool that can rightly account for two sentences with same meaning but different syntax can be hard, not to mention accounting for the narrative structure of plot or the roles of characters.

1. Narrative structures and markup languages

An interesting attempt to account for these structures in e-texts is Proppian Fairy Tale Markup Language (PftML) developed by Scott A. Malec. Based on Vladimir Propp's *Morphology of the Folktale* (1928), PftML utilizes a Document Type Definition (DTD) to create a formal model of the structure of Russian magic tale narrative and to help standardize the tags throughout the corpus. According to this approach, a text can be tagged as to encode Propp's "functions," or the 31 fundamental units that the Russian folklorist identified as the recurring building blocks of a Russian magic tale plot. Malec provides an example translated in English, The Swan-Geese tale. The related XML file is shown hereafter (see Listing 1), collapsed to 4 levels of depth (with the exception of the <Preparation> tag, which is fully expanded).

Software could easily parse this XML file, and quickly and reliably help the scholar in verifying some of Propp's theories, for example:

- **Acquisition of Magical Agent** appears three times, none of which directly lead to Victory,
- **Departure** is subsequent to **Villainy**,
- **Pursuit Rescue Of Hero** is composed by three series of **Pursuit Of Hero + Rescue Of Hero** (collapsed under <PursuitRescueOfHero> in the above picture)

```

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Folktale SYSTEM "http://clover.slavic.pitt.edu/~sam/propp/grammar.dtd">
<Folktale>
  <@Title>The Swan-Geese</@Title>
  <@AT>480</@AT>
  <@NewAdamskeEditionNumber>113</@NewAdamskeEditionNumber>
  <@ProppConfidence>Yes</@ProppConfidence>
  <@Source>
    <@Preparation>
      <@InitialSituation>
        Once upon a time a man and a woman lived with their daughter and small son.
        <@InitialSituation>
      <@CommandExecution>
        <@Command subtypes="Injunction">
          "Dearest daughter," said the mother, "we are going to work. Look after your brother! Don't go out of the yard, be a good girl, and we'll buy you a handkerchief."
        <@Command>
      <@Elevation subtypes="Violated">
        The father and mother went off to work, and the daughter soon enough forgot what they had told her. She put her little brother on the grass under a window and ran into the yard, where she played and got completely carried away having fun.
      <@Elevation>
      <@CommandExecution>
    <@Preparation>
    <@Villainy subtypes="Kidnapping"><@Villainy>
    <@ConsentToConstruction><@ConsentToConstruction>
    <@Departure>
      <@Leave subtypes="Departure">
        leaving all prepared quite arbitrary home --+
        <@Leave>
      <@DonorFunction subtypes="TestOfHero"><@DonorFunction>
      <@AcquisitionOfMagicalAgent subtypes="HelperOffersServices"><@AcquisitionOfMagicalAgent>
      <@DonorFunction subtypes="TestOfHero"><@DonorFunction>
      <@AcquisitionOfMagicalAgent subtypes="HelperOffersServices"><@AcquisitionOfMagicalAgent>
      <@DonorFunction subtypes="TestOfHero"><@DonorFunction>
      <@AcquisitionOfMagicalAgent subtypes="HelperOffersServices"><@AcquisitionOfMagicalAgent>
      <@Transferance subtypes="Runaway"/><@Transferance>
      <@StruggleVictory subtypes="Competition"/><@StruggleVictory>
      <@LiquidationOfLack subtypes="ReleasedFromCaptivity"/><@LiquidationOfLack>
      <@PursuitRescueOfHero><@PursuitRescueOfHero>
      <@Return><@Return>
    <@Monstrous>
  </Source>
</Folktale>

```

Listing 1: Example of XML Describing The Swan-Geese (Full Version: http://clover.slavic.pitt.edu/~sam/propp/have_a_little_byte/magicgeese.xml)

2. Graphic visualization of narrative structures

Taking Malec's work as a starting point, we produced a graphic representation of the XML file, aimed at highlighting some aspects of the narrative structure of the tale (see Fig. 1)

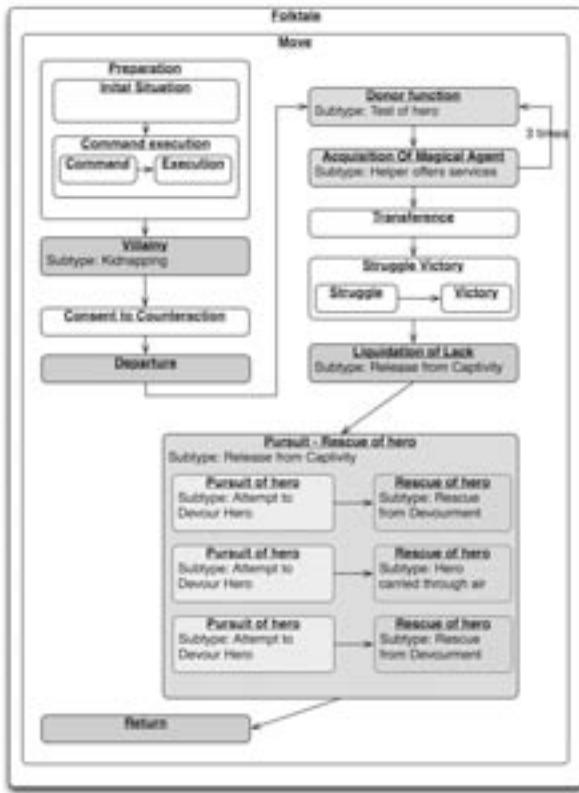


Fig. 1: Diagram Highlighting the Aspects of the Narrative Structure

The diagram above is meant to highlight the following aspects:

- Function nesting (e.g.: **Preparation** includes **Initial Situation + Command Execution**)
- Sequencing of “cornerstone” functions (**Villainy** -> **Departure** -> **Liquidation of Lack** -> **Return**).
- Cyclical repetition of same functions (in this case **Donor Function + Acquisition of Magical Agent**, and **Pursuit of Hero**).
- In case of repetition, some functions are of same subtypes (in this case the three instances of **Pursuit of Hero** are of same subtype, while **Rescue of Hero** instances are of two sub types, forming a A-B-A sequence).

3. The challenge

When we move from the analysis of words to the analysis of narrative structures we are facing a shift in the unit of analysis: we are no more limited to the elements of expression plane (i.e. words, in case of a text). We are now interested in “functions”, as named by Propp, or “events”, or “roles”. In other words, we are interested in analysing the meaning of sentences, or even entire paragraphs and chapters. In some cases the actual words used to express the meaning can be almost irrelevant for us, and we would like

to “see through” the endless possible variations of expressing the same concept.

In a folktale, if we are looking for that topic event generally called **Villainy**, regardless of whether carried out by villain or villain helper, which words should we look for? And if we are looking for **Liquidation of Lack**, represented in Sleeping Beauty by the re-acquisition of consciousness, can we consider these three sentences equivalent?

1. *The Prince kissed the Princess, and the Princess awoke.*
2. *The Princess awoke when kissed by the Prince.*
3. *The kiss given by the Prince awoke the Princess.*

We may say that these three sentences have different *phrase syntax* but same *actantial syntax*, as synthesized by Marsciani & Zinna (1991, pg. 56). That is, they express the same event (the Princess goes from asleep to conscious), with the characters having the same role (the Prince triggers the event), even if the words “Prince”, “Princess”, “kiss”, etc. appear to have different grammatical roles in each sentence. The word *actantial syntax* refers to Tesnière’s theory of *valency grammar*, where the verb is considered central to the sentence, like an atom that attracts a number of “participant roles”, the *actants* (Tesnière, 1959, p. 102). Tesnière explicitly attempts to analyse syntax and semantics separately (Tesnière, 1966 [1959], ch. 97 §3), and his work inspired the *actantial model* subsequently developed by A. J. Greimas (Marsciani & Zinna, 1991, pg. 54-57).

Using the notation proposed by Greimas, we could express the aforementioned

“the Princess goes from asleep to conscious” + “the Prince trigger the event”

by

$[S_1 \rightarrow (S_2 \cap O_1)]$	where S_1 = Subject 1, “the Prince” S_2 = Subject 2, “the Princess” O_1 = Object of Value 1, “consciousness” \cap = Union \rightarrow = Action
------------------------------------	---

This is an over-simplification of Greimas’ actantial model. However, it is worth noting that Greimas’ model is meant to formalise not only the semantics of a sentence, but also the narrative structure of the whole text. On the one hand, Greimas’ model is heavily based on Propp theory (Schleifer 1987, pg. 121-126), on the other it goes far beyond the actantial model and seeks to analyse the path of meaning as it goes, in a given text, from deeper structures to surfacing structures, also known as *Generative Trajectory of Discourse* (Greimas & Courtés, 1979, pg. 157).

4. The model: the Generative Trajectory

In an attempt to formalise and graphically represent the narrative structure of e text, the Generative Trajectory proved to be a valuable starting point. This model is well suited for our purpose because:

- It is well-rooted into narratology.
- It seeks for a “*fundamental semantics and grammar*” of narrativity, focusing on the relationships between expression plane and content plane (meaning), as well as on different pertinence levels.
- It already offers some formalism to express and analyse meaning, inspired by structural linguistics.
- It is based on 40 years of semiotic studies and has already proved to be very effective in analysing an impressive variety of texts.

Encoding some elements of this semiotic model into e-texts in the form of tags allowed us to produce a prototype graphic representation of some narrative phenomena.

The elements of Greimas' theory that we took into account are:

- **Multi-level analysis:** signification is articulated in three different pertinence levels, from deep structures to surface structures (Marsciani & Zinna, 1991, pg. 132-133):
 - i. **SEMIO-NARRATIVE STRUCTURES:** tags marking *axiologies* and *modalization*, as such describing how deep values are positioned on the *semiotic square*, how these values orientate the *Narrative Programs*, how the *actants* take position within the Narrative Programs.
 - ii. **DISCOURSIVE STRUCTURES:** tags marking *thematization* and *figurativisation*, as such describing the actors, places and times that constitute the discourse.
 - iii. **TEXTUAL STRUCTURES:** the text itself.
- **Conversion across levels:** Tags at different levels are interrelated, and these relationships constitute the tangible aspect of the “conversion process” across levels, that is the Generative Trajectory going from more abstract (deeper) levels to more concrete (surface) levels, up to the manifest level: the text itself.
- **Narrative Programs Nesting:** Besides the Basic Narrative Program, the one that subsume the whole text, other sub-Programs are taken into account (Array of Narrative Programs, (Hebert, 2006)).

- Actantial Model.

- **Semiotic Square:** used to describe articulation and axiology of deep values across levels.

5. Conclusion

In order to give a visual representation to narrative structures of text, we need to formalise the semantic and syntax of these structures. To that end, we relied on the A. J. Greimas' Generative Trajectory theory. Implementing some aspects of this theory in an experimental mark-up language allowed us to generate graphic visualization of the underlying deep semantic structures of texts. Some other aspects of Generative Trajectory could be then implemented, for example the **Canonical Narrative Schema**, especially relevant for the analysis of folktales corpora.

References

- Greimas, A. J., Courtés** (1979). *Sémiotique, Dictionnaire raisonné de la théorie du langage*. Paris: Hachette.
- Tesnière, L.** (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck, 2nd ed. 1966.
- Schleifer, R.** (1987). *A. J. Greimas and the Nature of Meaning*. London: Croom Helm.
- Propp, V. Y.** (1928). *Morfologija skazki*. Leningrad: Academia, English translation: *Morphology of the Folktale*, The Hague: Mouton, 1958; Austin: University of Texas Press, 1968.
- Marsciani, F., Zinna, A.** (1991). *Elementi di Semiotica Generativa*. Bologna: Esculapio.
- Hébert, L.** (2006). 'The Actantial Model'. *Signo*. Louis Hébert (dir.) (ed.). <http://www.signosemio.com> (accessed 15 November 2009).

Standards, Specifications, and Paradigms for Customized Video Playback

McDonald, Jarom Lyle

jarom_mcdonald@byu.edu

Brigham Young University, USA

Melby, Alan K.

melbyak@yahoo.com

Brigham Young University, USA

Hendricks, Harold

harold_hendricks@byu.edu

Brigham Young University, USA

Culture is fully inundated with video--from the ubiquitous DVD and succeeding optical media formats, to locally stored digital bits passed between DVRs and video iPods, to over a billion Internet-streamed videos a day. Unfortunately, while those involved in humanities education and research know how widespread video usage is and are attempting to integrate such a rich medium into what they do, they are more often than not struggling, fighting against the medium and associated baggage rather than using video for their own purposes.

For all of the ways in which video differs from other forms of media, perhaps the most challenging obstacle to effectively utilizing video assets as objects of teaching and research is their inflexibility. Because of the complexity of video technologies and the pressure of external interests, video is an incredibly closed medium, especially when compared to text, image, or even audio. In many ways video resists fundamental activities of digital humanities inquiry such as metadata, structural, and segment analysis and annotation. What's more, video also is, technologically speaking, a linear medium; it is (as much as if not more than other media) architected to proceed continuously from point A to point B, serving up bits in order and only responding to very limited, legacy interface controls. Even the "interactivity" touted by content holders (such as DVD "extras") is a rigid, linear interactivity, designed to keep the control of playback under the stewardship and limited scope of the video producer rather than the needs of the learner, the desires of the scholar, or the tastes of the consumer. To encourage collaborative, reusable approaches to video (while avoiding legal pitfalls or isolationist tendencies that come with an extracted clip approach), we need to incorporate a more thorough, flexible, and widespread method of *customized* video playback.

The papers in this panel will focus on data-driven customized video playback (CVP), from theory and methodology to real-world use cases that are evolving and practical implementations that are both already in use as well as under development to meet the needs of the Humanities today. The first presentation will make the case for the fundamental groundwork for video asset analysis and eventual customized video playback, the Multimedia Content Description Interface (also known as MPEG-7). This XML standard for describing (both globally and in timecode-associated ways) video assets offers a markup solution that is complementary to common video encoding containers (such as MPEG-2 and MPEG-4), and, as XML, can be easily coupled with other relevant data and metadata standards as well. The second paper will present an argument for ways to take these video asset descriptions and use them to enable both people and technology to better facilitate customized video playback using a videoclip playlist specification (serializable as plain text, as XML, as JSON, or as any other data exchange format). With the segment descriptions of a thorough video asset description, a videoclip playlist can then define custom playback operations. The final presentation will demonstrate several use cases of customized video playback, along with working models for achieving the type of interactivity we desire with the technologies we have today, including demonstration of a CVP system in use at several university campuses.

The panel as a whole will seek to argue a unified justification and methodology of customized video playback, and invite future collaboration from the Digital Humanities community who can, if they desire, push these ideas further towards making our proposed standards, specifications, and paradigms as widespread, useful, and effective as possible.

PAPER 1

Finding the Best in Approaches to Video Asset Description

McDonald, Jarom Lyle

jarom_mcdonald@byu.edu

Brigham Young University, USA

From Google's Web Services to Wikipedia's DBpedia project to the underlying architecture of modern digital libraries, our notion of how to make data more semantic is moving (slowly but persistently) towards

ideal principles that the W3C lays out for what is commonly called "the Semantic Web." This is even true for the subject of my study, video data, albeit with much less of a semantically-inflected critical mass. There are a few solid, innovative investigations (such as the BBC's video portal and the many incarnations of the Joost video platform) that are or have been working to bring technologies such as metadata, RDF/RDFa, and SPARQL to the storage and dissemination of video (especially online video); but there is still a lot of work that needs to be done in order to make today's video assets truly useful in a way that Tim Berners-Lee would approve, a world "in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Scientific American).

While the Semantic Web includes a large number of topics too broad to cover in this proposal, I will focus on one particular aspect of semantic markup that does apply to video data. It is vital to underscore the unique nature of video as an object of *perception*--that is, video is meant to be played for a viewer with the linear, temporal nature in the forefront of experiencing the video. Thus to describe the data of a video asset, as a whole, in a useful way would necessarily require a structured analysis of more than just the metadata *about* the video that you might be able to achieve with Dublin Core, IEEE-LOM, or RDF; the most significant need is a system that can connect such semantic vocabularies to a thorough, analytic description of the video content itself in as close an approximation to the playback act as might be reasonably able to achieve--in other words, a workable time-coded markup language. This isn't to say that a video must be necessarily viewed chronologically; rather, given that video exists as bits served from time point A to time point B, it must be described that way in order to make use of the data encoded there. If a video asset has the right description of its segmented, time-coded parts (of which, naturally, there may be many versions based on who is doing the markup or who is using the materials), it will eventually allow for more than just watching the video; a segmentation model of video markup is essential for enabling a system of interactive, customized video playback.

Several options for such a language to use are available and have been somewhat explored both commercially and academically, but none are completely satisfactory. Naturally, given the success of the Text Encoding Initiative, it makes sense to consider its ability to function as a time-coded video markup system. In fact, Reside (2007) and Arneil and Newton (2009) have presented just such an idea at recent Digital Humanities conferences. The flexibility and thoroughness of the TEI makes it an attractive option; however,

while the speech transcription models can potentially provide time-coded descriptions of spoken elements of a video (and even be retrofitted to other elements of video content), because the TEI is a text-encoding framework, it lacks a temporal segmentation scheme designed specifically for existing models of video encoding and playback (for example, referring to multiple video or audio tracks, multiplexing metadata with the binary streams, etc.). Most projects exploring video markup descriptions also mention the W3C's Synchronized Multimedia Integration Language (SMIL). Since version 3.0, SMIL integrates a temporal segmentation model with one for spatial fragmentation, allowing semantic relationships both within and between video elements. What's more, SMIL is a W3C recommendation, offering the potential for tighter integration with web delivered video as it continues to mature. Several commercial endeavors (including the streaming platform Hulu) have incorporated SMIL into their playback process, allowing for a sophisticated combination of video annotation (for example, Hulu uses it for their advertisements and upcoming social viewing features) and search/retrieval (combining the time-coded markup with RDF metadata). Yet SMIL provides no scope for how to implement various temporal segment references, instead leaving that task up to playback mechanisms (of which there are currently very few for SMIL). More significantly, SMIL provides no way to refer to described segments out of the context of the document, making it difficult to design a URI scheme for accessing the various clips, something integral to having true semantic web functionality or building video players that can interpret the descriptions consistently. Video streaming servers such as those provided by Apple, Microsoft, and Adobe have all developed their own model for segmenting (either in markup or in actual bits) video data and serving it with instructions for playback, but in these cases the systems very heavily limit the metadata that can be included, and the resulting descriptions are completely coupled to the proprietary vendor technologies (for example, a set of Flash Streaming Media Server cue points is not portable to other systems without intervention).¹

For the past 10 years, the Moving Picture Experts Group has defined and refined what they've formally titled the "Multimedia Content Description Interface," also known as MPEG-7. MPEG-7, an XML-based, ISO/IEC standard, is an expansive, far-reaching specification that does many, many things (including defining itself and defining the language by which it defines itself and its various parts); what is of particular interest to this proposal is Part 9, "Profiles and Levels." Recognizing that there are many approaches and viewpoints surrounding video asset description (those mentioned above,

plus such systems at TV-Anytime, the SMPTE Metadata Dictionary, and even extensions to the Dublin Core standard), MPEG-7 seeks to be a superset of video markup, and the concept of "profiles" as laid out in Part 9 of the spec offers various focused schemas and methodologies that conform to the MPEG-7 spec but serve unique needs. Seeing a need for a general purpose, video-specific description language, Brigham Young University has collaborated with Motorola and the Japanese National Broadcasting Corporation to publish a "Core Description Profile" (CDP), a framework that utilizes MPEG-7 descriptions and provides all the necessary tools for time-coded, segmented, video annotation.

Every file that conforms to the CDP schema (a schema now included directly as part of the MPEG-7 specification and which has been released as open-source by ISO) must also conform to the MPEG-7 super-schema. In addition to the MPEG-7 root element and any necessary header information, a CDP document has a series of <description> elements that contain <MultimediaContent>; a simple example of such an element might look something like this:

```
<MultimediaContent xsi:type="VideoType">
<Video id="MainTitle">
<TemporalDecomposition>
<VideoSegment id="chapter1">
<TemporalDecomposition>
<VideoSegment id="chapter1scene1">
<TextAnnotation type="description">
<FreeTextAnnotation>opening credits;
music; village aerial view
</FreeTextAnnotation>
</TextAnnotation>
<MediaTime>
<MediaTimePoint>T00:00:00
</MediaTimePoint>
<MediaDuration>PT1M24S
</MediaDuration>
</MediaTime>
</VideoSegment>
<VideoSegment id="chapter1scene2">
<TextAnnotation type="description">
<!-- other types of annotations are
possible as well --&gt;
&lt;FreeTextAnnotation&gt;entering church;
bells; Count introduced
&lt;/FreeTextAnnotation&gt;
&lt;/TextAnnotation&gt;
&lt;MediaTime&gt;
&lt;MediaTimePoint&gt;T00:01:24
&lt;/MediaTimePoint&gt;
&lt;MediaDuration&gt;PT0M20S&lt;/MediaDuration&gt;
&lt;/MediaTime&gt;
&lt;/VideoSegment&gt;
<!-- remaining scenes go here --&gt;
&lt;/TemporalDecomposition&gt;
&lt;/VideoSegment&gt;
<!-- Remaining chapters go here --&gt;
&lt;/TemporalDecomposition&gt;
&lt;/Video&gt;
&lt;/MultimediaContent&gt;</pre>

```

Having been designed as a general-purpose video asset description schema, the CDP is the most promising format for defining video clip boundaries, including metadata and annotations. On the surface it may not seem much different from other markup schemas such as SMIL; however, the real power of the approach lies in combining a CDP-conformant description with other parts of the MPEG-7 specification. First of all, because MPEG-7 is the data description framework for the same group behind the MPEG-2 and MPEG-4 video encoding containers and codecs, MPEG-7 can be easily multiplexed directly into a video container. Additionally, as XML, a CDP video asset description can incorporate any other relevant data through namespacing, including TEI, Dublin Core, RDF relationships, or future information schemas. And finally, because it is a general purpose description framework, it can also be serialized into any needed format such as SMIL, IEEE-LOM (a Learning Object Metadata standard) or CMML (the Continuous Media Markup Language), an XML schema defined by the organization behind the Ogg media formats and promising tight integration with emerging HTML5 video technologies. To project this even further, imagine a robust, RDF-aware repository (such as FEDORA) full of digital video objects that connect video streams to valid MPEG-7 video asset descriptions, making videos easily discoverable, easily searchable, and ultimately, truly semantic. And finally, this approach to video asset description lays the foundation for, to return to Tim Berners-Lee's comment, enabling us to make video playback from computer to person customizable, flexible, and just what we need it to be.

References

- Berners-Lee, Tim, James Hendler and Ora Lassila.** *The Semantic Web*. *Scientific American* May, 2001. <http://www.scientificamerican.com/article.cfm?id=the-semantic-web>.
- Bush, Michael D., Alan K. Melby, Thor A. Anderson, Jeremy M. Browne, Merrill Hansen, and Ryan Corradini** (2004). 'Customized Video Playback: Standards for Content Modeling and Personalization'. *Educational Technology*. 44.4 : 5-13.
- MPEG-7 Overview.** <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>.
- W3C Semantic Web Frequently Asked Questions.** <http://www.w3.org/2001/sw/SW-FAQ#What1>.

Notes

1. It's notable to also mention MIT's Cross Media Annotation System (XMAS), a project developed over the last decade to incorporate time-aligned video annotation and commentary into Shakespeare classes at MIT. The approach that XMAS takes is much more in line with what we see as the proper way to annotate video; however, theirs is a closed system very tightly coupled to their specific Shakespeare needs, so it isn't known what technologies they're using or how portable those technologies might be.

PAPER 2**Videoclips as Playlists in Customized Video****Melby, Alan K.**

melbyak@yahoo.com

Brigham Young University, USA

The video interface that we are all familiar with has garnered universal acceptance. This is true both of the iconic symbols of playback control, as well as the actual functions of control that are allowed. Since the days of very early analog playback, media consumers have been allowed to do pretty much just the following:

- Insert (or open) media
- Play
- Pause
- Stop
- Fast-forward
- Rewind
- Volume control (including muting)

There have been a few additions to the list as media (especially video) technology has evolved; for example, with the introduction of laserdisc and subsequently DVD, "return to menu," "next/previous chapter," and "go to title" controls have entered the collective interface. Streaming media has also invoked the need for a "fullscreen toggle" control. And many playback systems are now allowing for the display of certain content-provider defined metadata or even a "scrub" bar for more flexible time placement. But such change comes slowly, and all of these playback controls are very simple, require a good deal of user intervention, and truly limit the video asset's effectiveness as a learning object.



Image 1: A current video playback interface, offering only standard, simple user controls (source: <http://commons.wikimedia.org>)

To provide an example of the limitations of commonly accepted playback, let me briefly describe what happened this semester in a Hebrew class at our university. The instructor wanted to use several clips from a commercial Israeli movie to help the students with their vocabulary. With current US laws and technological learning curves making it unreasonable to rip the clips from the DVD and edit them for her class's needs, she instead instructed them to check out the DVD from the reserve library, start playing it, jump to chapter 14, watch for 4 minutes and 17 seconds (rewinding if necessary to fully understand the dialogue), jump to 2 minutes into chapter 31, watch for 3 minutes, pause the DVD and go look at some of the resources she'd posted on her course website, return to the DVD, etc. etc. etc. This is an extremely ineffective way to use the video in a learning environment, but unfortunately, it's really the only methodology that is widely available right now.

Historically, this hasn't always been the case. Laserdisc technology provided unprecedented control for both users and instructional designers over the video asset, and a large number of rich learning experiences were created and shared on laserdisc. More recently, the concept of "WebDVD" seemed poised to recover some of the lost functionality when laserdisc didn't emerge outside of a niche market; but WebDVD didn't catch on, either. Technologies of streaming media (such as YouTube annotations or bookmark-driven systems in place at CNN, ABC, Hulu, and other commercial streaming media institutions) have some potential, but they are still in their infancy, aren't widely used, and don't allow for anything other than an editorial overlay; the content itself is still played back under very strict control that the user can only pause, stop, rewind, etc.

What we propose, then, and what is so desperately needed, is a completely different playback system that allows for true interactivity for instructional

designers, teachers, and viewers. Bush et al. propose two models for customizing video playback that seek to alleviate the sort of haphazard "playback list" illustrated earlier. One, of course, would be to strip the digital bits onto a local filesystem, edit the content as needed, add in annotations (subtitles, links, external info), and share the new video with all students to view in traditional playback systems, repeating the process when the video content might need to be viewed in a different way. But this (as the authors point out) is time consuming and expensive, not to mention the unfortunate copyright implications of such an approach.

The other model for enabling customized video playback, and what we are currently developing, is to combine a data-driven, "descriptive" approach with a "selective playback mechanism", software specifically built for customized video playback according to a robust, standardized specification for delineating the different actions that the designer/instructor might want to have students experience. The most promising form of video description is the Video Asset Description (VAD), an XML encoding of clip boundaries, video content, and other metadata that is isomorphic with the MPEG-7, part 9 core description profile and which is described in another paper on this panel. When a video asset is associated with a full VAD (or several of them), an instructor--or in many cases even automated software--can use that description to generate a playlist. When I say "playlist," I'm not referring to the common usage of the term as a description of a media collection (a list which describes what assets to play and in what order), but it is similar; what I propose is a notion of a video clip playlist (VCP), a description of timecoded clips within a video asset. A collection of these clips, along with the actions to take for each clip, could be fed directly into the queue of selective playback software, software that would be programmed to know how to read the instructions and present the new playback session.¹

Each instruction in a videoclip playlist would be a triple that would consist of a framecode number, an operation, and an operand. The framecode is not a direct representation of either human perception of time or of frame count, but is instead a convenient fiction that allows for the most effective and standardized accuracy in calculating either time or frame. The operation, most easily represented as a numeric opcode, would be one of a number of operations that a playback system might encounter. Of course there would be the standard "play, pause, stop, mute" controls, but they would be under the stewardship of the instructional designer who is authoring the videoclip playlist. There would also be codes for jumping to a new timecode (not just a chapter, not just an approximate location on a scrub

bar, but a frame-accurate location), for displaying annotations (subtitles, scholar-composed notations, instructor comments, etc.), displaying "wrap" data (for example, material retrieved from web services and displayed in an additional pane of the selective viewer at the precise moment the playlist instructs), and so forth. The operand would be a piece of data that makes the op code intelligible; if a command instructed the player to jump to a new time code, the operand would be the time code to jump to. If a command instructed the player to start playing a clip, the operand might be the number of frames to play.

We have designed a simple RNG schema for encoding these clips in an XML file that is both machine and human readable. The file would have some header information that identifies the videoclip playlist, associated video asset descriptions, and references to wrap data and other annotations, as well as an instructionList of the commands that the player would need to perform the custom playback. The instructionList looks like this:

```

<instructionList>
    <instruction trigger="0" opCode="68"
    operand="60">show 'skipping' message for 2
    seconds</instruction>
    <instruction trigger="0" opCode="0"
    operand="0">pause before seeking</instruction>
        <instruction trigger="0" opCode="65"
    operand="52164">seek to frame 52164</
    instruction>
        <instruction trigger="52164" opCode="75"
    operand="32">new clip [32] begins</instruction>
            <instruction trigger="52164" opCode="85"
    operand="0">show wrapData #0</instruction>
                <instruction trigger="52164" opCode="85"
    operand="1">show wrapData #1</instruction>
                    <instruction trigger="52164" opCode="85"
    operand="2">show wrapData #2</instruction>
                        <instruction trigger="53753" opCode="68"
    operand="60">show 'skipping' message for 2
    seconds</instruction>
                        <instruction trigger="53753" opCode="0"
    operand="0">pause before seeking</instruction>
                            <instruction trigger="53753" opCode="65"
    operand="99926">seek to frame 99926</
    instruction>
                            <instruction trigger="99926" opCode="75"
    operand="57">new clip [57] begins</instruction>
                                <instruction trigger="99926" opCode="85"
    operand="3">show wrapData #3 </instruction>
                                    <instruction trigger="99926" opCode="85"
    operand="4">show wrapData #4 </instruction>
                                        <instruction trigger="99926" opCode="85"
    operand="5">show wrapData #5 </instruction>
                                            <instruction trigger="100911" opCode="75"
    operand="58">new clip [58] begins</instruction>
                                                <instruction trigger="100911" opCode="85"
    operand="6">show wrapData #6 </instruction>
                                                    <instruction trigger="100911" opCode="85"
    operand="7">show wrapData #7</instruction>
                                                        <instruction trigger="100911" opCode="85"
    operand="8">show wrapData #8 </instruction>
                                                            <instruction trigger="103033" opCode="99"
    operand="-1">end of playlist -- indefinite
    pause

```

```
</instruction>  
</instructionList>
```

A player, of course, would only need to be passed the triples represented by the integer values of each instruction's attributes, and in fact a videoclip playlist could be serialized in any necessary data exchange format, whether it be JSON (for building a browser-based player for customizing streaming media playback), plain text (that might include just tab-delimited integers easily consumable by an appliance with an embedded selective player), and so forth.

Obviously, one key to such an approach to facilitating customized video playback is the creation of the selective players themselves. We are currently undergoing development on specifications that would allow anyone to build such a player. With the combination of robust video asset descriptions, shareable, thorough videoclip playlists, and intelligent, VCP-aware players, customized video playback is once again a reality.

Notes

1. Some might ask why it's necessary to go to the trouble of having a markup layer associated with time-aligned video segments at all; why not use extracted clips? However, the legal and technical obstacles involved in extracting segments of video are far greater problems than those experienced through the copy/paste of small snippets of text, making clip extraction unfeasible for most cases (not to mention the fact that clips themselves are just as rigid, and must be re-extracted if the use cases change). Moreover, extracted clips are difficult to share and collaborate on, and would still need some sort of annotation layer associated with them for editorial commentary, additional subtitles, etc. Our proposed methodology can handle annotation, collaboration, modification, re-use, and legal restrictions all with one approach.

PAPER 3

Customized Video Playback; Where We've Come From, Where We're Going

Hendricks, Harold

harold_hendricks@byu.edu

Brigham Young University, USA

When we talk about customized video playback, it's important to recognize that the actual playing back of the video is paramount; theories of how to

mark up the video's content or describe the desired playback are significant only insofar as they can lead to actual implementations that satisfy some of the use cases that we might envision for a customizable video playback system. These use cases generally fall into three types: one-on-one interactivity, classroom lecture, and large audience presentations, such as annotated cinema. I hope to move the locus of attention from CVP theory to practice in three ways: by discussing some of our historical attempts to achieve such an implementation, by demonstrating a current, working system for customized video playback in use in several academic institutions today, and by outlining where our work is moving next (and how we envision collaborating with others outside our project who have so much to contribute).

The introduction of videodisc technology in the mid 1970s provided the first practical method for inexpensive video storage and random-access playback, especially in the realm of academic instruction. *Macario*, a repurposed Mexican motion picture, was issued as a custom videodisc pressing with interactive menus and annotations coded to the linear playback of the video. When a student would pause the video, the interactive materials would appear, allowing for commentary, instruction, thought-questions tied directly to individual scenes being watched, and even replay of selected video with choice of audio track.

Macario's model of annotated video playback offered some innovative learning opportunities to, for example, intermediate Spanish language and culture classes, but it was still a fairly simplistic model, one built upon and improved over the next few years. Projects such as the German Video Enhanced Learning, Video Enhanced Teaching (VELVET) program empowered students with more custom tools, such as the ability to filter out or select particular types of annotations (both text and image), highlight keywords in accompanying transcriptions, or even perform intricate searching through accompanying materials to narrow in on particular scenes (replaying them as needed) of use to the student. These types of activities demonstrated how useful customized video playback could be, focusing the viewing experience and tailoring it to particular educational needs.

In the late 1980s, Junius Bennion and Glen Probst modified some of these previous models of interactive video to allow more control of video assets within targeted learning experiences. Having first created a methodology for an Apple II-controlled videodisc of *Raiders of the Lost Ark*, Bennion, Probst and James Taylor reprogrammed the content to work with Hazeltine's Time-shared, Interactive, Computer-Controlled, Information Television (TICCIT) System

at BYU. Within this model the motion picture is divided into scenes with light-pen interactivity with annotations, transcriptions, questions, text and audio commentary, and instructional drills. Examples of these TICCIT programs include versions of *Black Orpheus*, *The Seventh Seal*, and *C'eravamo tanto amati*. With the TICCIT modified model, customized, interactive video instruction moved from research projects to the language lab.

The ideas underlying some of these models of customized video playback are the same principles expounded upon in the other sessions of this panel, implemented in the best possible way using the technology available at the time. But they were all inextricably linked to the technology itself, needing custom produced videodiscs or a complex networked computer system to run. When videodisc technology never caught on (for a number of reasons), these innovative products were made obsolete. Likewise our attempts to achieve robust customized video playback through such frameworks as HyperCard, ToolBook, and "WebDVD" have struggled for much the same reason. Recently, however, we have been able to achieve quite functional and effective CVP through an existing system entitled "Electronic Film Review" (EFR), a methodology for controlling video playback of DVDs.

The EFR approach, demonstrated as a poster session at the 2006 Digital Humanities Conference, is based on the MPEG-7 and VideoClip Playlist open standards discussed earlier, and is designed to be implementable in any media player for time-coded video that supports playing a segment of video based on time codes. The current implementation of the EFR approach runs on Windows XP computers that have decoders suitable for watching DVDs through shared, DirectShow DVD decoders (the current EFR software does not include its own DVD decoders). For individual language study, each user-defined clip of a film can be annotated with vocabulary, culture, and other notes. The EFR player itself includes the video window, the common media interface controls, custom "playlist controls" (for navigating between pre-defined clips), and areas to display the annotations. The EFR system also includes an authoring tool, EFR Aid, as well as a compiler to generate the playlist format, to ease the definition and annotation of various segments of a particular video.

Because the EFR system is based on open standards, any learning materials created for particular video assets are shareable; video asset descriptions and playlists can be transferred from one user to another. What's more, these resources that the EFR system helps create are not coupled to the video asset itself; they are a form of meta-annotation (hence the title of "film review") that do not interfere with a single bit

of the video data, thus respecting any copyright laws that might exist. As plain-text (serializable as XML), they are also fully searchable, allowing for discovery of relationships between videos that may not have been previously known. Most importantly, the EFR system makes video much more than just watchable; it makes video useable.



Image 2: A screenshot of the EFR video player

By useable, once again we mention the three primary use cases: individual interactivity, classroom interactions, and annotated cinema. Both our historical efforts and the current EFR implementation have focused primarily on a single user interacting with a computer, with some efforts made to enhance classroom presentation. However, the EFR program has successfully been used in all three of these cases, providing the means for enhanced comprehension, vocabulary building, speech modelling, and cultural awareness for individuals, a means to integrate these same video-based activities into the classroom, and also a tool for modifying the playback of full-length feature films with content filtering and additional subtitles without modifying the copyrighted and encrypted video.

Now, as mentioned earlier, several universities (Brigham Young University, the University of Hawaii at Manoa) have used or are currently using a Windows-XP based EFR system for DVD playback. However, once again the evolution of technology is forcing change in our approach to customized video playback. With the introduction of Blu-Ray, the explosive growth of online, streaming video, and constant legal and political fighting between content providers and content consumers, we see it necessary to broaden the scope of the EFR project to allow for all possible use-cases that we might imagine. We are currently undergoing a project, a collaboration between academic institutions and commercial enterprises, to define open specifications for building a CVP system for any technology. Our goal is to create content and meta-data that can

be used on any machine without worry that the necessary technology might not be available. Though still in their nascent stages, these specifications will build on the principles outlined throughout this session--reusable, robust XML markup of video assets, clip divisions, annotations, and playback instructions. If we are successful, we hope to end up with standards that can be used regardless of the video encoding format or delivery system. We invite suggestions and participation from the community as we move forward from the historical and current availability of customized video playback towards an approach that works for all time-based media now and in the future.

References

- Hendricks, Harold.** (1993). 'Models of Interactive Videodisc Development'. *CALICO Journal*. **11.1**.
- Melby, Alan** (2004). 'The EFR (Electronic Film Review) Approach to Using Video in Education'. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2004*. L. Cantoni & C. McLoughlin (ed.). Chesapeake, VA: AACE, pp. 593-597.

The Origins and Current State of Digitization of Humanities in Japan

Muller, A. Charles

acmuller@jj.em-net.ne.jp
University of Tokyo

Hachimura, Kōzaburō

Ritsumeikan University

Hara, Shoichiro

Kyoto University

Ogiso, Toshinobu

National Institute for Japanese Language and Linguistics

Aida, Mitsuru

National Institute for Japanese Literature

Yasuoka, Koichi

Kyoto University

Akama, Ryo

Ritsumeikan University

Shimoda, Masahiro

University of Tokyo

Tabata, Tomoji

University of Osaka

Nagasaki, Kiyonori

International Institute for Digital Humanities

Digital Humanities in Japan has been in progress since an early period. Recently, due to the spread and the development of advanced digital environments, individual humanities researchers are coming to use digital materials in various forms, and according to the continually growing needs of users, cooperation and organization between projects has steadily increased. However, there is some extent to which the large framework known Humanities Computing in Japan has lagged behind in its efforts to develop cooperation with similar projects overseas. Therefore, this panel aims to take a step in the right direction by introducing the origins and current state of Humanities Computing in Japan on digitization by featuring the reflections of the representatives of the projects and organizations that have worked in this area from a relatively early time.

Humanities digitization projects in Japan, being developed by various research centers and research organizations, have already garnered over 3.5 billion

yen in the form of known large-scale grants which were funded by Japanese Ministry of Education, Culture, Sports, Science and Technology. This does not include the numerous small grants that have been received for various digitization projects by individual researchers and small groups and budgets which were assigned by each organization itself which was promoting such projects. In order to provide a venue for the presentation and publication of the results of these funded projects, a number of groups and associations engaged in this work were established. One of the more prominent is Special Interest Group for Computers and Humanities (SIG-CH), which was established in 1989 under the auspices of the Information Processing Society of Japan which is the largest society of informatics in Japan. SIG-CH has served as the major organ for communication among researchers interested in these projects.

PAPER 1

Kōzaburō Hachimura, Ritsumeikan University, representing SIG-CH

The meetings of SIG-CH have been held on a regular basis, about four times a year since 1989. At each meeting approximately 8 research papers are presented, with the proceedings being published as the "IPSJ SIG Notes" series. Up to the present, we have held 84 meetings, which have included a total of 720 paper presentations. The group consists of over 200 researchers belonging to the academic organizations of informatics or humanities.

Here, processing, analysis or mining of texts, images, digital archiving of texts, bibliographies or other digitized materials, and especially 3-D motion capture, etc. are the major themes. The technique, tools or study results in which they are applied etc. are presented. In recent years, study results that use GIS have been increasing in number. The various Humanities fields in Japan are represented by literature, linguistics, history, archeology, museum studies, anthropology, dance studies, and Buddhist Studies, etc. In earlier periods, system-oriented thought was dominant, but recently, presentations tend to be characterized by an increase in content orientation, as well as on local and international cooperation between projects.

PAPER 2

Toshinobu Ogiso, representing the National Institute for Japanese Language and Linguistics NINJAL)

NINJAL, established in 1948, has created various Japanese corpora. One example is the "*Taiyō Corpus*," which is the first major Humanities database in Japan created using XML. It is a tag-structure rendition of the 19-20th century magazine *Taiyō*, in which tags created for the purpose of linguistic research applied to about 14.5 million text characters. *Taiyō* was a typical magazine in Japan during the period of its publication from 1895 to 1925, and thus is an invaluable resource for understanding the foundations of the modern Japanese language which were formed during that period.

Also underway at this institute is the KOTONOHA plan, which seeks to integrate various corpora (including the *Taiyō* corpus). One part of this effort is the presently-underway project of the 'Balanced Contemporary Corpus of Written Japanese' (BCCWJ), containing 100 million words. In addition to this, the construction of a corpus is planned aimed at compiling premodern data. This project must especially address the peculiarities of written Japanese, which does not include spaces between words, and includes Chinese characters, *hiragana*, and *katakana*, making it very difficult to indicate word information with pauses between phrases, parts of speech, etc.

PAPER 3

Mitsuru Aida, representing the National Institute for Japanese Literature (NIJL)

NIJL was established in 1972, making one of the earliest efforts to digitize Japanese literature. Researchers there have worked at converting the research information into database format, and in its inclusion of words, text descriptions, and literary indexes, has become Japan's prototype textual research database. In the early 1990's, NIJL defined an original standard for tagging Japanese literatures based on SGML, and upon this built a large full-text database. It has played a major role as the mechanism for a general database of the human culture research.

PAPER 4**Koichi Yasuoka, Kyoto University, Institute for Humanities Research (Jinbunken)**

In 1980 the Jinbunken began the digitization of the *Ming Dynasty Civil Examination Index*. The following year, the institute initiated the digitization of the *Index of Shanwen Liyi* and *Catalog for the Study of East Asian Documents*. Moreover, the Institute has held an Annual Workshop for Oriental Studies Computing (ORICOM) every year in 1990. The research conducted here for the past 20 years extends to many areas, including multilingual text processing, character-code issues, digital catalogs, and GIS. Most notably, the Jinbunken has been actively engaged in critique and development of Japanese *kanji* character sets, dealing with issues concerning the relationship between ancient characters, JIS X 0213, Unicode and so forth.

PAPER 5**Ryo Akama, representing the Ritsumeikan University Art Research Center (ARC)**

The extensive works of ARC have their origins in the digitization of the *kabuki* material that Prof. Akama initiated in the Waseda University Theater Museum in 1988. This approach led to the establishment of the ARC in 1998. The ARC has worked on the digitization of various material and intangible Japanese cultural treasures through the aid of the Ministry of Education. The research is carried out making good use of various information technology skills, which include not only text and images, but also 3D images and motion capture, etc. At present, they have received a Global COE grant from the Ministry of Education, whereby they are serving as a base for Digital Humanities intended for Japanese culture as a whole.

PAPER 6**Masahiro Shimoda, representing The University of Tokyo Center for Evolving Humanities (CEH)**

One of the major aims of the CEH is, along with the development of its own Buddhist texts information system, to bring about cooperation with a wide range of digital projects related to the study of Indian Philosophy and Buddhism, and in so doing, to demonstrate a solid example of the possibilities of digital humanities studies to the Japanese academic world. The project includes the development of an extensive bibliographical database for the field of Buddhist Studies in Japan, which now includes about 70,000 entries. In addition, the Daizōkyō Text Database Research Committee, established in 1994, has completed and released a set of highly accurate text data of 600MB, covering the major portion of the East Asian Buddhist canon. This is known as the SAT Database, which is now fully interactive with the above mentioned article database, as well as the online reference work, the *Digital Dictionary of Buddhism*. The project is working toward further expansion of cooperation with other data bases in the field.

PAPER 7**Conclusion**

The projects represented here constitute only a very small portion of what is going on in Japan. For example, on the educational front, the Faculty of Culture and Information Science of Doshisha University, established in 2005, is aiming to teach methods of research for the analysis of cultural information at both the undergraduate and graduate levels, and in the future is expected to be a major source for the development of talented researchers in the digitization of humanities. A few decades have now passed for Humanities Computing in Japan, and a wealth of data, tools, and techniques have been produced as a result. We are convinced that this panel will be an important step for locating digitization of humanities in Japan in the flow of worldwide Digital Humanities, and in charting our own future.

Born Digital: The 21st Century Archive in Practice and Theory

Gabriela Redwine

gredwine@mail.utexas.edu

Harry Ransom Center, The University of Texas at Austin

Matthew Kirschenbaum

mkirschenbaum@gmail.com

University of Maryland

Michael Olson

mgolson@stanford.edu

Stanford University Libraries / Academic Information Resources Stanford University

Erika Farr

elfarr@emory.edu

Robert W. Woodruff Library, Emory University

As more people rely on computer technologies to conduct their personal and professional lives, born-digital materials such as emails, Word manuscripts with tracked changes, blog entries, text messages, and tweets will constitute the archives of the future. Archival repositories at places like Stanford University, Emory University, and The University of Texas at Austin have been receiving born-digital materials for over 20 years but have only recently begun working actively to preserve these items in their original digital formats.

As part of this work, archivists have begun to look to other fields, such as computer forensics and law enforcement, for equipment and methodologies to use in the acquisition and preservation of born-digital materials. The application of forensics technology to born-digital content in archives and the development of tools to facilitate access to these materials hold great promise for humanities scholarship and teaching.

This session brings together digital archivists, librarians, and curators to discuss some of the forensic techniques and equipment being used to preserve born-digital archival materials at the Stanford University Libraries, the researcher interfaces Emory University has developed to provide access to Salman Rushdie's computers, and the broader implications of these developments for the concept of "archives" in a variety of disciplines, including information science, literary studies, history, and cultural studies.

Michael Olson, Digital Collections Project Manager for Stanford University Libraries, will begin the session with a discussion of the applicability of forensics software to the acquisition and description of born-digital archival materials at Stanford. Erika Farr, Director of Born-Digital Initiatives at Emory's Woodruff Library, will discuss the researcher interfaces developed for use with Salman Rushdie's computers and the results of user studies currently underway to explore the potential effects of analog-digital hybrid materials on research methodologies and scholarly communication. Gabriela Redwine, Archivist and Electronic Records/Metadata Specialist at the Harry Ransom Center, will consider the computer as an archival object that challenges both archival and scholarly notions of what an archives is and can be, as well as the functions it may serve.

The panel will be chaired by Gabriela Redwine, of the Harry Ransom Center, The University of Texas at Austin. Matthew Kirschenbaum, Associate Director of the Maryland Institute for Technology in the Humanities (MITH), will serve as respondent.

PAPER 1

Computer Forensics in the Archive: An Analysis of Software Tools for Born Digital Collections

Michael Olson

mgolson@stanford.edu

Stanford University Libraries / Academic Information Resources Stanford University

Stanford University Libraries hold an increasing amount of digital archival material. This principally comprises magnetic and optical disks and tapes containing digital files produced both via historical computing platforms on legacy media, as well as via contemporary applications on modern media. Analysis of recent acquisitions from the last five years has shown a five-fold increase in the number of collections containing digital archival materials. Without near-term action, these materials are at the greatest risk of loss and are likely to disappear from the corpus of primary source materials. The imminent loss of digital archival materials now confronts curators, digital archivists, and researchers who desire to use and preserve these digital records.

Computing forensics is a discipline that is still very much dominated by the law enforcement community and the need for digital evidence that can be verified in a court of law. It is based on the following core principles: "that evidence should not be altered, examination results should be accurate, and that examination results are verifiable and repeatable" (Pollitt, 1995). These same principles translate to the archival world, where provenance or verifiable custody is a foundation of archival theory. Curators, digital archivists and researchers have the same requirement that documents, whether in an analogue or digital format, be verifiable.

Digital investigations, both criminal and commercial, have driven the development of forensic software tools and training. Commercially produced forensic software and training certification programs are almost universally adopted by law enforcement agencies. Open-source software for the capture and analysis of digital archival materials is available as an alternative, but there is even less data on how these tools work or could be used in the archival field.

Beginning in early 2009, staff from Stanford's Digital Libraries Systems and Services group met with our archivists to assess the preservation and access needs for digital archival materials. Out of these discussions at-risk collections were identified and it was determined that our highest priority was to safely migrate these collections off at-risk media in a forensically sound manner. Our greatest concern was that the floppy disks, magnetic tapes, and hard drives in our collections would degrade before we could develop a comprehensive program to both preserve and make these materials available to researchers. A second priority was to acquire software tools that would allow our archivists to assess the contents of digital materials and develop methods for making them available.

Alongside this priority-setting exercise, Stanford sought advice from the participants at the British Library's Digital Lives Conference. Jeremy Leighton John at the British Library and staff from the Paradigm Project (co-directed by Oxford and Manchester) were particularly helpful in providing their expertise and a list of potentially useful hardware and software (Paradigm, 2005-7). Following up on this advice Stanford began an intensive discussion with multiple forensic vendors that currently supply and train many law enforcement agencies in the United States. These discussions were notable by the surprise many forensic firms expressed when presented with our archival needs; law enforcement is clearly driving the market for forensic hardware and software.

In the summer of 2009, Stanford University Libraries acquired a suite of forensic hardware and software

and has undertaken an extensive program to test a wide range of commercial and open-source forensic software applications and evaluate which applications are most appropriate for use by our curatorial staff, digital archivist, and donors. This paper summarizes our experience in evaluating our academic archiving needs against the range of commercial and open-source forensic software applications. It is important to note that our findings are not scientific product evaluations. The results provided in this paper merely reflect our own experience using these different methodologies to forensically image and analyze digital archival materials from the perspective of a curator, a digital archivist, and a potential donor of digital archival materials.

The results of our findings are based on the following criteria: the nature of the archival collection, skills required to use the software effectively, an evaluation of feature sets, potential for integration with existing archival software such as the Archivists' Toolkit, support for metadata outputs and preservation services, application cost, and supported forensic disk image formats. Our non-scientific evaluation includes two of the largest commercial applications used by the forensic law enforcement community: Guidance Software's EnCase Forensic™ and AccessData's FTK (Forensic Toolkit) 3.0™. In addition, we will include our evaluation of open source software such as The Sleuth Kit and a small number of freely available forensic utilities.

References

- Paradigm project** (2005-7). *A Proposal for Intellectual Access to Hybrid Archives. Workbook on Digital Private Papers.* <http://www.paradigm.ac.uk/workbook/cataloguing/intellectual-access.html> (accessed 12 November 2009).
- Pollitt, M. M.** (1995). 'Principles, Practices, and Procedures: An Approach to Standards in Computer Forensics'. *Second International Conference on Computer Evidence*. Baltimore, Maryland, 10-15 April 1995.

PAPER 2

Finding Aids and File Directories: Researching a 21st Century Archive

Erika Farr

elfarr@emory.edu

Robert W. Woodruff Library, Emory University

The introduction of desktop computers, MD5 checksums, handheld devices, and digital forensics into archives and special collections brings with it a transformation of accessioning procedures, processing practices, preservation tactics, and research service approaches. The impacts of these shifts and transformations will be felt not only by archivists and librarians but also by researchers and scholars.

In this paper, I will discuss how the arrival of born-digital content into archives has insisted on innovations in archival practice and promises to bring significant change to research methodologies. As a practical, concrete means of framing this discussion, I will focus on a particular case study: Salman Rushdie's hybrid "papers," housed in Emory University's Manuscript, Archives, and Rare Book Library (MARBL). By considering the acquisition, processing, and accessibility of this collection, this paper will discuss the new challenges introduced to archival science by such hybrid collections. More importantly for the purposes of this paper, user testing and user studies currently underway on the Rushdie materials will provide valuable data and insight into how hybrid collections of primary materials may influence archival research habits and scholarly communication.

The 2006 acquisition of Salman Rushdie's papers, which included both traditional manuscript materials and a series of personal computers, provided Emory University Libraries with its first significant hybrid collection of personal papers. With the exception of a few articles (e.g. Thomas and Martin, 2006) and the *Workbook on Digital Private Papers* produced by the Paradigm project (2005-7), very little documentation existed to guide the staff at MARBL and in Emory's Woodruff Library in its approach to accessioning and handling these materials. Early in the development of the Rushdie project and Emory's Born-Digital Archives program, the team made a commitment to approach the material as holistically as possible, prioritize the integration of paper and digital, and balance donor requests with researcher needs. Such a philosophy

prompted us to begin processing by first capturing complete disk images of all five hard disks, then creating verifiable MD5 checksums, and revisiting security and confidentiality concerns at virtually every processing turn. Our comprehensive interest in the collection demands that our development of access points and tools embrace both the digital context (e.g. the operating system, original applications, original file formats) and the larger context of the complete collection (e.g. paper materials and finding aids). This interest in context led us to explore virtualized environments as a point of access and resulted in the development of researcher tools that allow concurrent exploration of emulated environments, the finding aid, and item-level, database-driven searches.¹

In addition to providing a greater level of detail about the early processing of and planning for the Rushdie papers, this paper will also highlight important early collaborations. In particular, I will discuss some of the valuable insights gained while participating in an NEH Office of Digital Humanities Start-up Grant with partners from the Maryland Institute for Technology in the Humanities and from the Harry Ransom Center at the University of Texas at Austin (see Kirschenbaum et al., 2009). This start-up grant has had important influences on our program development.

As this planning grant was concluding, Emory began finalizing plans for the public release of Rushdie's archive. In preparation for this major milestone in February 2010, staff at MARBL and in the Digital Systems division of Emory's Woodruff Library worked diligently to process the materials in both traditional and more innovative ways and to create tools, infrastructure, and interfaces that will enable effective researcher access to a selection of the born-digital materials as well as the finding aid for the paper materials. With a completed prototype of the researcher workstation ready for initial testing in early October 2009, staff undertook a cornerstone piece of work for the Born-Digital Archives program: user testing. Because born-digital archival content changes how researchers access and interact with materials, it will necessarily result in changes in how researchers undertake their work. In order to provide optimal support of and service for such scholarly pursuits, archives and libraries must relentlessly explore, study, and analyze researcher needs and habits. Given the current transformative period in archives, it is especially important that we know, even anticipate, what researchers will want to do with these materials ten, twenty, even fifty years from now.

In an essay discussing researcher habits in archives, Duff and Johnson argue that archives need more accurate and diverse scenarios of use in order to

better understand how scholars use and interact with archival material (2002, p.473).² This observed need for more data and better understanding about how researchers currently use archival materials fuels Emory's interest in gathering user feedback on born-digital materials and exploring effective interfaces for such collections. With this mission at the program's core, we will continue to undertake testing and user studies, beginning in earnest in March 2010. Based on findings and results from these studies, we will take the practical steps of revising and augmenting our systems and services, as well as undertaking the slightly more theoretical activity of documenting these habits in order to begin articulating shifting methodologies in scholarly research.

This paper will elaborate on the activity and development of Emory's Born-Digital Archives program, expound on the work involved in providing researcher access to Rushdie's hybrid collection, and introduce early findings from user studies and testing on the initial set of tools produced for the release of Rushdie's hybrid archive. Discussion of these activities within the framework of how hybrid collections impact research and supported by data gathered during studies and testing should begin to illuminate some of the ways in which research may evolve and transform in the twenty-first century archive.

References

- Duff, W. M., Johnson, C. A.** (2002). 'Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives'. *Library Quarterly*. **72(4)**: 472.
- Kirschenbaum, M.** (2007). 'Hamlet.doc?: Literature in a Digital Age'. *The Chronicle of Higher Education*. **53(50)**: B8-9. <http://chronicle.com/free/v53/i50/50b00801.htm> (accessed 20 October 2009).
- Kirschenbaum, M. et al** (2009). *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use*. NEH Office of Digital Humanities. <http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37> (accessed 2 November 2009).
- Paradigm project** (2005-7). 'A Proposal for Intellectual Access to Hybrid Archives'. *Workbook on Digital Private Papers*. <http://www.paradigm.ac.uk/workbook/cataloguing/intellectual-access.html> (accessed 30 October 2009).

- Thomas, S., Martin, J.** (2006). 'Using the Papers of Contemporary British Politicians as a Testbed

for the Preservation of Digital Personal Archives'. *Journal of the Society of Archivists*. **27(1)**: 29-56. <http://doi:10.1080/00039810600691254>.

Notes

1. In his *Chronicle of Higher Education* article "Hamlet.doc?: Literature in a Digital Age," Matthew Kirschenbaum's description of the rich potential of born-digital papers demonstrates one example of scholarly interest in born-digital material beyond discreet files.
2. Duff and Johnson focus on historians in this piece, but their conclusions and observations pertain to humanities research more broadly.

PAPER 3

Archives and 'the Archive': The Computer as Archival Object

Gabriela Redwine

gredwine@mail.utexas.edu

Harry Ransom Center, The University of Texas at Austin

In 2009, the National Endowment for the Humanities funded a project entitled "Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use," which supported site visits among personnel working with the born-digital components of three significant collections of literary materials: the Salman Rushdie Papers at Emory University, the Michael Joyce Papers at the Harry Ransom Center, and the Deena Larsen Collection at the Maryland Institute for Technology in the Humanities (MITH). In the two publications emerging from that project, the grant collaborators, and Matthew Kirschenbaum in particular, articulated the idea of an author's computer as what Kirschenbaum termed a "complete material and creative environment"—one that an author inhabits like she would a suit of clothes, or an office, or a self (Kirschenbaum et al., 2009a and b). This paper will build on that understanding of the relationship between computer and author to consider the ways in which the computer, as a complex archival object, pushes the boundaries of traditional archival practice and also has the potential to reshape the discussion of "the archive" as a subject of critical inquiry.

The forensic techniques Stanford, Emory, the Ransom Center, and other repositories are using to capture images of disks and hard drives offer the potential for archivists to preserve and analyze

more information about authors' work and lives than ever before. For example, an author's browsing history could provide insight into her online research during a particular period of creativity, or the trash folder of an email account could contain discarded emails important to an understanding of a particular manuscript. The tools being developed as part of forensic projects like Simson Garfinkel's Real Data Corpus can be used to recover data and characterize relationships between data sets. For example, it would be possible to map social networks between computers and create a visualization showing which authors were communicating with each other during a certain period of time. This type of work could be done using hard drives residing at a single repository or as part of a collaborative project across institutions. But does the existence of these types of materials and the technological capability to preserve and analyze them mean that archivists should? What is potentially hidden or revealed when a laptop or a server-based Twitter or email account is part of a collection acquired by an archival repository? What ethical concerns arise around born-digital manuscript drafts deleted by a creator, files "hidden" within a computing system, or correspondence that exists only in the cloud?

The Society of American Archivists, North America's oldest professional association for archivists, defines an archives as a body of "materials created or received by a person, family, or organization, public or private, in the conduct of their affairs and preserved because of their enduring value" (2005). Questions of value have long been at the center of debates among archivists, scholars, activists, historians, politicians, governments, and others about what gets saved, by whom, and to what end. The implications of historical definitions of archives and the presumably objective role of the archivist continue to inform scholarship in a variety of fields. One of the most influential examples is *Archive Fever* (1996), in which Jacques Derrida challenges the concept of an archive as a definable entity with estimable value and an uncomplicated relationship to history and memory. In the seminal essay collection *Refiguring the Archive* (Hamilton et al., 2002), contributors ranging from Derrida to Verne Harris (Nelson Mandela's archivist) to Achille Mbembe (historian and postcolonial theorist) debate the relationship of archives to memory in the context of South Africa's social, cultural, and political history. Archivists such as Michelle Light and Tom Hyry have argued for greater transparency on the part of archivists and an acknowledgement of the subjectivity inherent in organizational and descriptive practices (2002). And scholars like Ann Cvetkovich have articulated a broader understanding of the concept of an "archive," beyond the types of records and other materials found in conventional

archives, to include non-traditional, and often more ephemeral, representations of things like memories, feelings, and lived experiences. *Writing in An Archive of Feelings* (2003) about cultural spaces constructed around sex, feeling, and trauma, Cvetkovich laments that their "lack of a conventional archive so often makes them seem not to exist."

So how might forensic technology and its affordances impact the ways in which creators, archivists, and scholars perceive what information is buried or accorded cultural value and whether and how to describe it? How might computers, as complete material and creative environments, make it possible for an individual (or a group) to generate an archive that preserves the ephemeral, the transformative, the everyday, the personal, the painful, and much more, in a variety of audio, visual, and textual genres? My exploration of these and other questions will incorporate the work of the cultural and literary theorists mentioned above, as well as more traditional archival texts and definitions. This paper will consider the ways in which the computer as an archival object challenges notions about the role of archives, the concept of the "archive," and the work of archivists in global contemporary cultures. I will pay particular attention to the ways in which global disparities in access to technology risk creating a future archive that in many ways resembles the colonial archive of the past.

References

- (2005). 'Archives'. *A Glossary of Archival and Records Terminology*. Society of American Archivists. <http://www.archivists.org/glossary/> (accessed 9 November 2009).
- Cvetkovich, A.** (2003). *An Archive of Feelings: Trauma, Sexuality, and Lesbian Public Cultures*. Durham: Duke University Press.
- Derrida, J.** (1996). *Archive Fever: A Freudian Impression*. Chicago: University of Chicago Press.
- Hamilton, C., Harris, V., Taylor, J., Pickover, M., Reid, G., Saleh, R. (eds.)** (2002). *Refiguring the Archive*. Cape Town: David Philip Publishers.
- Kirschenbaum, M., et al.** (2009a). 'Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use'. NEH Office of Digital Humanities. <http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37> (accessed 4 March 2010).
- Kirschenbaum, M., et al.** (2009b). *Digital Materiality: Preserving Access to Computers as Complete Environments*. San Francisco, CA, 5-6 October 2009.

Light, M., Hyry, T. (2002). 'Colophons and annotations: New directions for the finding aid'. *American Archivist*. **65(2)**: 216-230.

Networks of Stories, Structures and Digital Humanities

Almila Akdag Salah

almila.akdagsalah@vks.knaw.nl

KNAW Royal Netherlands Academy of Arts and Sciences, Netherlands

Wouter De Nooy

w.denooy@uva.nl

UVA, Faculty of Social and Behavioral Sciences, Netherlands

Zoe Borovsky

zoe@ats.ucla.edu

UCLA, ATS, LA, USA

In March 2004, Slashdot, the online news aggregator 'for nerds', cited a web page featuring animated network visualizations of the relationships between characters in Shakespeare's plays.¹ The Shakespeare site began getting hits at a rate of 250,000 hits per hour. Programmer Paul Mutton created the diagrams by feeding Shakespeare's plays into an IRC (Internet Relay Chat) bot designed to visualize social networks. The Slashdot story promised that the diagrams would allow users to "see Shakespeare in an entirely new light".²

Social network analysis, the mapping of relationships as networks, has truly provided new insights in the social and life sciences with increasing participation from mathematicians and computer scientists. The idea that there are laws that govern networks--such as the notion that everyone is at most six steps away from any other person on earth, and that networks evolve in predictable ways--has led to remarkable discoveries in fields from sociology to biology. These insights furthered publications such as Six Degrees (by sociologist Duncan Watts)³ presenting these discoveries (and the science behind them) to a more general audience. In addition, online communities such as Facebook, Twitter, MySpace, etc., have demonstrated the utility and power of network theory in our daily lives.

With popular awareness of these tools and techniques, combined with the promise of new insights into ever-increasing amounts of data, network analysis has significant appeal to digital humanities scholars. However, these investigations call for tools and software to gather and clean large amounts of data, even when the researcher opts for analyzing only a tiny fraction of these vast data

sets. Moreover, to interpret these datasets, a good understanding of network analysis, as well as other visualization means is necessary. Overcoming the hurdle of handling complicated technological tools and acquisition of the necessary expertise in network theory both demand investment on the side of researchers, a risky investment that only a handful of humanities scholars are willing to make.⁴ Typically, researchers familiar with the tools perform these analyses, posing the questions and interpreting the results. In this session, we will first provide a brief overview of social network analysis and related tools. Then, we will focus on three examples of how we, as humanities scholars, have made use of network analysis tools and techniques in our research, both to illustrate the potential of these approaches, and to discuss some common problems and their possible solutions.

The names and affiliation of confirmed authors are as follows:

- Zoe Borovsky, Academic Technology Services and the Center for Digital Humanities, University of California, Los Angeles
- Wouter de Nooy, Amsterdam School of Communication Research (ASCoR), University of Amsterdam, The Netherlands
- Loet Leydesdorff, Amsterdam School of Communication Research (ASCoR), University of Amsterdam, The Netherlands
- Andrea Scharnhorst, Virtual Knowledge Studio of the Netherlands Royal Academy of Arts and Sciences, The Netherlands
- Almila Akdag Salah, Virtual Knowledge Studio of the Netherlands Royal Academy of Arts and Sciences, The Netherlands

Notes

1. <http://www.jibble.org/shakespeare/>
2. http://tech.slashdot.org/story/04/03/11/0151256/Tracking-Social-Networking-In-Shakespeare-Plays?art_pos=24 Mutton also presented his work at a conference on Information Visualization in 2004. See <http://www.cs.kent.ac.uk/pubs/2004/1931/content.pdf>
3. Watts, D.J. (2003). *Six Degrees: The Science of a Connected Age* 1st ed., W. W. Norton & Company
4. Social network analysis is a more popular tool among social scientist than humanities scholars. A search at Web of Science for the search term returns around 150 hits for social sciences, whereas if the search is refined to arts and humanities only, it has less than 30 results.

PAPER 1

Dickens' Double Narrative: Network Analysis of Bleak House

Zoe Borovsky

zoe@ats.ucla.edu

Academic Technology Services and the Center for Digital Humanities, University of California, Los Angeles

My presentation extends Masahiro Hori's collocational analysis of Dickens' style (2004)¹ using network content analysis tools. Because DH2010 will be held in London, and since Hori devotes a chapter in his book to Dickens' novel, *Bleak House*, I will focus the presentation on this novel. My goals are three-fold: First, I will show how network analysis provides additional insights and metrics for studying the collocational patterns in texts. Secondly, I will demonstrate how this analysis provides new insights into the gender issues in Dickens' novel. Finally, I will extend my analysis to the field of Digital Humanities and reflect on how it, too, can be read as a "double-narrative" with a deep structure similar to the one Dickens depicts in *Bleak House*.

Network Analysis has been applied to many different types of systems: sociological, biological, ecological, as well as the Internet. In addition, human language has been examined for evidence of network behavior—the laws that have been shown to govern other complex systems. Beginning with the 2001 publication of "The Small World of Human Language".² I will present a brief overview of scholarship on network analysis to provide some background. Typically, humanities scholars have used network analysis to model networks described in texts (e.g. the characters in Shakespeare's plays).³ However, the main purpose of this paper is to demonstrate the implications of these more theoretical studies upon text analysis.

Having provided the context for my analysis, I will show how Wordij,⁴ a suite of network content analysis tools designed to analyze unstructured texts, can be used to extend Hori's analysis of the collocates⁵ of the two narrators in Dickens' text. Against the prevailing view that the first-person narrative of the central female character, Esther, is plain and boring, Hori compares the unusual collocates of both narrators. His careful analysis reveals that Esther's narrative is "linguistically experimental and satisfactorily creative" (Hori 2004,

p. 206). My analysis builds upon Hori's and illustrates that through the course of the novel, the narrative of the anonymous third-person narrator becomes "tainted" with collocates that have led scholars to judge Esther's narrative as overly sentimental and boring.

Because Wordij allows us to analyze texts as a network, with links created between collocates, we can use it to calculate the distance between concepts that are connected through other links. Using the network diagrams produced from collocates of both these narrative sections, I will demonstrate how this type of "semantic map" can be used to compare the distance between words. As examples, I compare the distance between the words "man" and "woman" in each narrator's section. Opticomm (one of the programs in Wordij) measures the distance between nodes (the words) by counting the number of links between words and calculating the distance or path and the average pair frequency.

For example, one of the shortest paths from the word "man" to "woman" in Esther's text takes the following route: man→said→woman. Opticomm calculates the "distance" using the frequency of the pairs. The pair "man→said" occurs 27 times ($1/27 = .0370$), and the pair "said→woman" occurs 19 times ($1/19 = .0526$). The sum of the distances equals .0897. The average pair frequency of the path is low: .23. A higher frequency path takes advantage of more frequently used pairs. One of the "hubs" (a word that has many connections) in Esther's narrative is the word "little". By adding the word "little" to the path (man→said→little→woman), the distance between "man" and "woman" becomes shorter (.0673) and more frequent (53.6667) [See Figure 1].

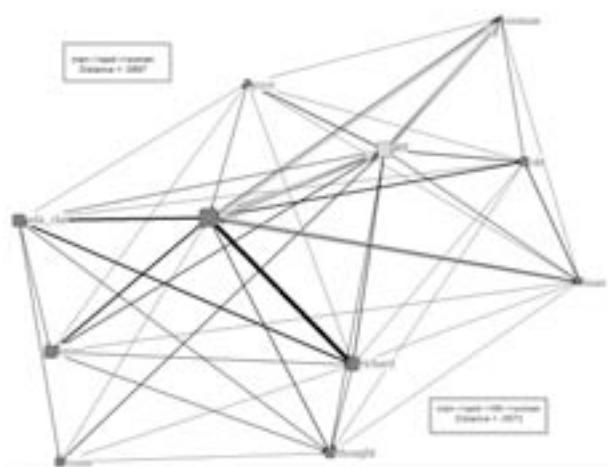


Figure 1: Network of collocates of Esther's narrative, Dickens, Bleak House. Visualization is created with NetDraw 2.087 using collocates and distances generated with Wordij. Frequencies of the pairs (shown by weighted lines) are greater than or equal to 18.

The results of this analysis show how the third-person narrator's change in style (employing

collocate patterns used in Esther's text) reflects and mirrors the character development in the novel—the rezoning of the hyper-masculine characters: Sir Leicester Dedlock and Mr. George. I will argue that the deep structure of the novel hinges on this "rezoning" of the masculine— a plot structure that is remarkably similar to the occult horror films analyzed by Carol Clover in her book: *Men, Women and Chainsaws: Gender in the Modern Horror Film*.⁶ Like occult films, I will present that the aim of Bleak House is a reconstruction of the masculine, a process that typically takes place almost invisibly, as the texts foregrounds the excesses of the female characters. Thus the narrative technique, one that "opens up" to the sentimental expressions characteristic of Esther's narrative, reflects the deep-structure "rezoning" on the periphery of Bleak House; George's Shooting Gallery and Sir Leicester's Chesney Wold are radically transformed in the novel.

Finally, I will draw comparisons between the cultural conditions that evoke these types of structures and double narratives, and reflect on how the field of Digital Humanities can be read in this light. Scholarship in Digital Humanities is a "double-narrative" brought about by the rezoning of hard and soft sciences, a project similar to the reconstruction undertaken by Dickens' *Bleak House*.

Notes

1. Hori, M. (2004). Investigating Dickens' style, Palgrave Macmillan.
2. Cancho, R.F.I. & Solé, R.V. (2001). The Small World of Human Language. Proceedings of the Royal Society of London. Series b, biological sciences, 268, 2261--2266.
3. Stiller, J., Nettle, D. & Dunbar, R. (2003). The small world of Shakespeare's plays. Human Nature, 14(4): 397-408.
4. Danowski, J., WORDij, Chicago: University of Illinois at Chicago. Available at: <http://wordij.net/>
5. Collocates, or word-pairs, have been defined and analyzed in various ways. Hori provides a useful short history in his book. Wordij generates collocates or pairs of words that occur within a user-specified span of surrounding text. See Danowski, J. (2009). Network analysis of message content. In Krippendorff, K & Bock, M. (eds), The content analysis reader. Sage Publications, pp. 421-430 for a detailed description of how Wordij generates and ranks collocates.
6. Clover, C.J. (1993). Men, Women, and Chain Saws: Gender in the Modern Horror Film. Princeton University Press.

PAPER 2**Network analysis of story structure****Wouter de Nooy**

w.denooy@uva.nl

Amsterdam School of Communication Research
(ASCoR), University of Amsterdam, The
Netherlands

Relations and structure play an important role in the humanities. Language is structured at the word, sentence, text levels, and beyond. Literary texts contain implicit references to one another and artists are linked by lines of influence or artistic descent. In some branches of structuralism, meaning is inextricably intertwined with the relations among linguistic elements (signifiers). In this paper, I extend the idea that structure is the carrier of meaning to stories. I submit that story structure conveys meaning. Patterns of interactions among the characters of a story are meaningful in the sense that they represent roles with moral connotations.

Structure quickly becomes complex if the number of elements and relations increases. Then, network analysis is an indispensable tool. In a network, entities such as characters in a story are represented by nodes (vertices) and lines embody links between these characters. There is no limitation to the kinds of links that can be represented by lines as long as each link connects two entities. Durable ties, for example, family relations among characters, can be expressed in this way just as easily as incidental events or interactions. Lines can be symmetric, for example, being siblings, or asymmetric, such as an action from one character directed towards another character in the story. Lines can have attributes and they can be of different kinds, for example, positive versus negative lines (this is called a signed network). Finally, each line can be time-stamped to indicate the onset and expiration of ties. With this formalization of structure as a network, we can use computers to quickly and correctly identify structural patterns. Software that is able to analyze large and complicated networks is widely available.

If network analysis yields a tool for finding patterns, does it also tell us which patterns to look for? In an analysis of story structure understood as the interactions among characters, it makes sense to borrow from the tradition of social network analysis. In this paper, I use the theory of structural balance, which originated in (social) psychology and was adopted in social network analysis. Based on simple

premises like ‘the friend of my friend is my friend,’ this theory predicts that people favour balanced networks. A network is balanced if the positive (friendship) and negative (antagonism) ties display specific, easily recognizable patterns.

Balance theory has been applied to the affective ties among characters in stories: tales, opera, and movies. The general finding is that balance among the story’s protagonists is restored at a story’s resolution. Stories develop from unbalanced towards balanced networks. In this paper, I focus on another important role of time. I show that the time order and direction of positive and negative actions between two protagonists defines the roles they play in the story. I use the seminal work of Vladímir Propp—*Morphology of the Folktale* (1928)—which inspired most of the later work on story grammars and narratology. Propp offers a concise set of propositions on the basic form and plot of a corpus of Russian fairytales, distinguishing between several roles of protagonists, such as the hero, the villain, and the donor. I demonstrate that each role is associated with a unique sequence of positive and negative actions. In other words, one can tell a fairytale character’s role from the interaction sequence it is involved in.

A formal focus on structure, which is needed for computerized network analysis, entails that attributes of the characters, such as a hut on chicken legs indicating that a character is a witch, and detailed content of interactions are not taken into account. Instead, the focus is on an abstract schema that can be fleshed out in any number of ways in a story. Notwithstanding their abstract character, I contend that the schemata have moral implications. Attacking the story’s hero out of the blue, which is a characteristic of the structural pattern associated with the villain, has a negative moral connotation, which is also linked to the last step in the villain’s structural pattern, namely that the villain is defeated by the hero at the end of the story. Undesirable properties of the villain, such as ugliness, may help to reinforce the moral depreciation of the role but they are not essential to it.

Tales and other stories socialize children into a culture of appreciating some sequences of interaction and depreciating other sequences. Authors may play with these conventions by combining structural role patterns with attributes of the character that are usually associated with other roles. In a more fundamental way, authors may turn structural patterns upside down, for example, a woman accepting to live with the man who raped her (a typical villain role) in Coetzee’s *Disgrace* or the protagonist in Houellebecq’s *Elementary Particles*, who turns his back on his most important ‘donor.’ Fairytale structure is probably quite straightforward

in comparison to stories read by adults. A systematic analysis of more complicated stories requires formal network analysis and so does the detection of conventional role patterns in sets of stories, differences in this regard between story genres, and development over time.

If the role patterns in stories are instruments of socialization, we may expect to find them in real-life interactions. Because role patterns and network analysis are purely structural, it is straightforward to compare the structure of interactions within stories to interactions among humans. In this paper, I will apply network analysis to evaluations among a set of literary authors and critics to show that fairytale roles can be distinguished and help to explain what happens. In this way, the social sciences benefit from the analysis of story structure just like social theory and social network analysis may advance the analysis of story structure.

PAPER 3

Mapping the Flow of “Digital Humanities”

Alkim Almila Akdag Salah

almila.akdagsalah@knaw.vks.nl

Virtual Knowledge Studio of the Netherlands Royal Academy of Arts and Sciences, The Netherlands

Loet Leydesdorff

loet@leydesdorff.net

Amsterdam School of Communications Research (ASCoR), University of Amsterdam, The Netherlands

Andrea Scharnhorst

andrea.scharnhorst@vks.knaw.nl

Virtual Knowledge Studio of the Netherlands Royal Academy of Arts and Sciences, The Netherlands

In this paper, we propose novel ways of charting out activity in digital humanities research. In particular, we apply a combination of classic bibliometric analysis¹ focusing on citation patterns and a broader text and actor network analysis based on various data sets collected from different sources, such as standardized scientific information databases and web data in general.

"Digital humanities" is the latest term for "humanities computing", almost a century old enterprise that made use of computational methods and tools in humanities research. Today, this enterprise has become a hot topic supported

by significant government funding agencies like National Endowment in Humanities (NEH) in the United States² and the European Framework programs of EU.³ The topic attracts attention from scholars coming from a wide range of disciplines like computer science and art history, and its future is the focus of a heated debate on certain weblogs dedicated to this research initiative. Depending on the research environment, the idea of Digital Humanities (DH) is envisioned in diverse ways; some research institutes or universities view it as a methodology, others incorporate it into their research agendas as a tool for analysis. Others view it as a discourse, investigating how knowledge is produced with new media technologies, while at the same time making use of these technologies in humanities research itself.

Usually, the applications and methodology of DH ask for a collaborative commission, and these interdisciplinary collaborations forge beneficial links that enable the use of technological tools in a way to serve humanities research questions. However, recent findings (Leydesdorff & Akdag Salah, in preparation)⁴ show that the scholarly activities around DH are less broad than one could expect. One might even say that DH falls short of reaching out to an interdisciplinary audience. A bibliometric study of the citation patterns of papers that are tagged in Web of Science's all three databases under the topic of "digital humanities" reveals that the knowledge base of DH is confined to two groups of journals, namely the ones that are dedicated to humanities computing, and the others dealing with library and information sciences (see Figures 1 & 2). This outcome should not be surprising, as research in the arena of humanities computing was centered upon text mining and analysis of humanities literature, and the publication of these research results were confined to specialized journals such as *Humanities Computing*. Beside text analysis, a main focus of DH activities was the digitization of archives and libraries, hence the second group of journals are specialized in this area.

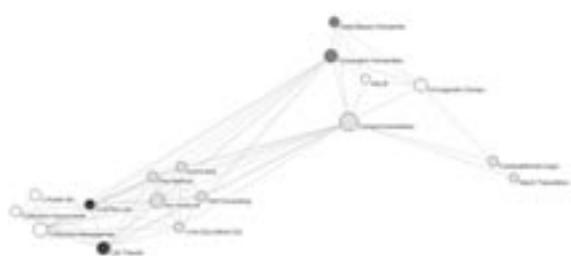


Figure 1: Journal co-citation patterns of 33 documents citing 46 documents about "Digital Humanities" in 2008; threshold 0.5%; cosine ≥ 0.0 ; N = 15, from Leydesdorff & Akdag Salah (forthcoming).



Figure 2: k-core group of 36 journals bibliographically coupled in 46 documents about “digital humanities”; threshold 0.5%; cosine > 0.2, from Leydesdorff & Akdag Salah (forthcoming).

Currently, it is not clear whether DH is in the first formations of either becoming a discipline on its own, or whether it will become merged into the infrastructure of future humanities departments by offering a platform for collaborative research environment enriched by media technologies of today. In either case, our initial findings on DH's position and its present inability to reach an interdisciplinary audience led us to conclude that policy makers and scholars, both of whom contribute to the development of DH could benefit from an expansion of our analysis.

We assume that DH activities that are very innovative and still in an emergent state might not be fully reflected by the journals sampled in *Web of Science*. We furthermore suppose that a more in-depth analysis that goes beyond journal-journal citation networks will reveal interesting patterns. We will present the results of a social network analysis and a semantic map of DH based on a dataset extracted from papers published on leading DH journals such as *Humanities Computing*, *Calica*, *Literary and Linguistic Computing*, as well as papers retrieved from online publication venues such as *Text Technology*, *Digital Studies*, *Digital Medievalist*. We believe that online journals will alter the recent findings of the knowledge dissemination of DH to a considerable extend. Here one should not be limited with classical database providers like *Web of Science* or *Scopus*, but also make full use of Google Scholar search-engine. Even then, a bibliometric study of the topic confined to journal-journal citation maps will not deliver a complete picture of the flux of DH. To capture this change and the debate around it, we propose a methodology that will incorporate different type of social networks in rendering and interpreting this new research area.

DH scholars can use virtual research environments for collaboration and production of knowledge, as well as for publication of their results. Hence, information retrieval from online communities, mailing lists, discussion forums and blogs dedicated to DH research is crucial for understanding the state

of the art in projects, programs, and research done in the name of DH. A social network based on the shared keywords of such a dataset reveals what topics are of recent interest to DH scholars. Furthermore, this dataset is used in preparing a second network that shows the links of websites and blogs (in and out), which facilitates the drawing of information flow inside DH community, as well the community's relation to other scholarly communities.

A very interesting extension is a map that depicts the connection between institutions and the resources they draw upon. As we have already mentioned, the structural formation of DH is different in every institution. Thus, every DH research environment is built on different premises, and the development of each of these settings is mainly based on internal resources (which are usually delivered in terms of equipment or the availability of technical staff and scholars) and on external support (governmental or private agencies). Today, when the main attention of DH scholars are on combining their past experience and their ideas about DH to create a better platform for humanities departments in general, to depict the flow of resources, financial and otherwise, is as vital as following the flow of information in DH.

Notes

1. ‘Scientometrics’ or ‘bibliometrics’ is a research venue specialized in evaluating growth, relations and interactions in scientific fields with the help of citation data collected above all from journal papers. Scientometrics is traced back to the beginnings of the 1900s, but the more official start can be settled to 1964, when Garfield founded Institute for Scientific Information; today it is a product of Thompson ISI and called Web of Science. From the 80’s onwards, the research in this area accelerated with the advancement of computers and various combinations of statistical methods used to extract and evaluate information such as citations, cocitations with reference of various bibliometric data. The end-results are usually rendered as so-called ‘citation networks’ which are a variation of social networks. Now it is a common practice to evaluate a scholar or a journal according to how many times it/he/she is cited. To read more on the history of scientometrics see Katy Börner, Jeegar T. Marus, and Robert L. Goldstone. (2004). The Simultaneous Evolution Of Author And Paper Networks, PNAS 101 (suppl.1): 5266-5273. For more information on citation networks, please see Doreian Patrick. (1985). A Measure Of Standing Of Journals in Stratified Networks. Journal of the American Society for Information Science, 8(5/6): 341-363. For a critique of bibliometric analysis see Lindsey D. (1989). Using Citation Counts As A Measure Of Quality In Science Measuring What's Measurable Rather Than What's Valid., Scientometrics, Vol. 15, No 3-4: 189-203; Leydesdorff L. (2006). Can Scientific Journals Be Classified in Terms of Aggregated Journal-Journal Citation Relations Using the Journal Citation Reports? Journal Of The American Society For Information Science And Technology—March: 601-614.
2. NEH has an office dedicated for Digital Humanities research: <http://www.neh.gov/whoweare/divisions/DigitalHumanities/index.html>

3. One of the FP7 projects is titled PREPARINGDARIAH (Preparing for the construction of the Digital Research Infrastructure for the Arts and Humanities) and is devoted to enriching DH: <http://www.dariah.eu>
4. Leydesdorff L., Akdag Salah, A. (forthcoming). Maps on the basis of the Arts & Humanities Citation Index: the journals Leonardo and Art Journal, and “Digital Humanities” as a topic, Journal of the American Society for Information Science and Technology, available at <http://www.leydesdorff.net/ahci/ahci.pdf>

Understanding the 'Capacity' of the Digital Humanities: The Canadian Experience, Generalised

Siemens, Ray

siemens@uvic.ca
University of Victoria, Canada

Eberle-Sinatra, Michael

michael.eberle.sinatra@umontreal.ca
Université de Montréal, Canada

Siemens, Lynne

siemensl@uvic.ca
University of Victoria, Canada

Sinclair, Stéfan

sgsinclair@gmail.com
McMaster University, Canada

Brown, Susan

sbrown@uoguelph.ca
University of Guelph, Canada

Timney, Meagan

mbtimney.etcl@gmail.com
University of Victoria, Canada

Rockwell, Geoffrey

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

Chair: Ray Siemens

Presenters (4-5 minute presentations):

- Michael Eberle-Sinatra, “Understanding Academic Capacity: A Charge from our Funding Agency”
- Lynne Siemens, “Developing Academic Capacity in Digital Humanities: Thoughts from the Canadian Community and Beyond”
- Ray Siemens, “DH Training Capacity: Established Curriculum, Institutes, Camps, and Beyond”
- Stéfan Sinclair, “Building from the Ground up: Training Digital Humanities Scholars as Developers”
- Susan Brown, “Hidden Capacity (in DH-impacted disciplines)”
- Meagan Timney, “Transitions: Emerging in the Field”

- Geoffrey Rockwell, "Cyberinfrastructure for Research in the Humanities: Expectations and Capacity"

Panel Description: In a recent report from Council of Canadian Academies, *The State of Science & Technology in Canada* (2006, p. 24; <http://www.scienceadvice.ca/documents/Complete%20Report.pdf>) humanities computing was identified as an "emerging field" with "significant strength," alongside several other science-oriented "transdisciplinary fields ... for which future prospects are seen to be more significant than currently established strength." Concomitant discussions with our chief research funding agency, the Social Science and Humanities Research Council of Canada (SSHRC), yielded a need to understand, better, the 'capacity' of this community in Canada and beyond, in part to gauge the potential future impact of our interdiscipline which, itself, has been generously supported for a decade with dedicated programs such as SSHRC's Image, Text, Sound Technology program and research infrastructural programs such as the Canadian Foundation for Innovation (CFI). Manifest in a report commissioned by SSHRC, our panel discusses the results of and reactions to the activity of attempting to understand our field's capacity, both within the national context that spawned the study and the borderless environment occupied by the field.

The panel consists of several very short presentations, followed by discussion. **Michael Eberle-Sinatra**'s presentation will discuss specific aspects of the report itself, among them the field's history of interrelationship with supporting programs, societies, and initiatives; how the field presents itself to, and intersects with, the larger humanities community; the field's notable successes, notable contributions, and chief projects; and what is anticipated to be needed next to enable excellent and timely research across the humanities, from the perspective of the field – all through the lens of the fields enabling possibilities via methods, tools and cyberinfrastructure. **Lynne Siemens** will juxtapose the growing acceptance of digital humanities research and teaching methods, technologies and resources with a series of challenges that still face scholars, especially new scholars, in developing their work in the field, via the results of a recent survey of digitally-impacted faculty, staff and students in the Humanities and Social Sciences (which yield a focus on funding, infrastructure and leadership). **Ray Siemens** discusses the crucial role of training, broadly construed, a point which will be picked up on and carried much further by **Stéfan Sinclair**, who acknowledges that the training of humanists with advanced programming skills is essential to the digital humanities' recognition of tool

conception and development as first class scholarly activities, to the process of building as a way of exploring and understanding, touching also on important issues of peer-review and professional recognition for innovative work in tool-building. **Meagan Timney**, a postdoctoral fellow in digital humanities, discusses several issues specifically confronting emerging scholars in the field of digital humanities and **Susan Brown** will posit, from the position of someone at a university that has no formal digital humanities programs or even dedicated courses in the calendar, that there is considerable untapped capacity for digital humanities training in the Canadian, and other, higher education systems – highlighting the fact that there are many people with significant training or research experience in the digital humanities area teaching across the humanities where their work has as a matter of course included the impact of digital textuality but they would not identify with our field by its name, the result of underfunding of the traditional disciplines and lack of institutional resources to create new programs or organise existing offerings in new, pertinent configurations. **Geoffrey Rockwell** will close by asking "what is infrastructure in the humanities?" – presenting a model for the research computing infrastructure we should expect from our universities and suggesting the capacity at Canadian universities to meet this need as well as the politics of positioning computing as infrastructure.

Coalition of Humanities and Arts Infrastructures and Networks - CHAIN

Wynne, Martin

martin.wynne@oucs.ox.ac.uk

University of Oxford, UK

Anderson, Sheila

sheila.anderson@kcl.ac.uk

King's College London (DARIAH)

Fraistat, Neil

fraistat@mac.com

Maryland Institute for Technology in the
Humanities, University of Maryland (centerNet)

Kainz, Chad

cjkainz@uchicago.edu

University of Chicago (Project Bamboo)

Krauwer, Steven

s.krauwer@uu.nl

Utrecht University (CLARIN)

Robey, David

d.j.b.robey@reading.ac.uk

University of Oxford (Network of Expert Centres)

Short, Harold

harold.short@kcl.ac.uk

King's College, London (ADHO)

A panel will discuss areas of cooperation and practical work between the various initiatives engaged in building e-infrastructure to support the next generation digital research in the Humanities.

Following from a panel session at DH2009, in October 2009 representatives of numerous important associations, networks and projects met and resolved to form *CHAIN* – the Coalition of Humanities and Arts Infrastructures and Networks. The initial 'CHAIN Gang' comprised:

- arts-humanities.net
- ADHO - Association of Digital Humanities Organisations
- CLARIN
- centerNet
- DARIAH
- NoC - Network of Expert Centres in Great Britain and Ireland
- Project Bamboo

- TextGrid

The representatives of these various initiatives identified the current fragmented environment where researchers operate in separate areas with often mutually incompatible technologies as a barrier to fully exploiting the transformative role that these technologies can potentially play. *CHAIN* recognised that their current and planned activities were interdependent and complementary and resolved that they should be oriented towards working together to overcome barriers, and to create a shared environment where technology services can interoperate and be sustained, thus enabling new forms of research in the Humanities.

CHAIN will act as a forum for areas of shared interest to its participants, including:

- advocacy for an improved digital research infrastructure for the Humanities;
- development of sustainable business models;
- promotion of technical interoperability of resources, tools and services;
- promotion of good practice and relevant technical standards;
- development of a shared service infrastructure;
- coordinating approaches to legal and ethical issues;
- interactions with other relevant computing infrastructure initiatives;
- widening the geographical scope of our coalition.

CHAIN will promote an open culture where experiences, including successes and failures, can be shared and discussed, in order to support and promote the use of digital technologies in research in the Humanities.

This session will feature panellists from the *CHAIN* gang, who will introduce the practical measures that they are engaged in to build a coherent, interoperable and maximally effective services to support research in the Humanities, and will address the following questions:

- What are the main barriers to progress?
- What are the most exciting opportunities?

Papers

Character Encoding and Digital Humanities in 2010 – An Insider's View

Anderson, Deborah

dwanders@sonic.net

UC Berkeley, USA

The world of character encoding in 2010 has changed significantly since TEI began in 1987, thanks to the development and adoption of Unicode (/ISO/IEC 10646) as the international character encoding standard for the electronic transmission of text. In December 2008, Unicode overtook all other encodings on the Web, surpassing ASCII and the Western European encodings (Davis 2009). As a result, Unicode's position seems to be increasingly well-established, at least on the Web, and TEI was prescient to advocate its use.

Over 100,000 characters are now defined with Unicode 5.2, including significant Latin additions for medievalists, a core set of Egyptian hieroglyphs, and characters for over 75 scripts. As such, Unicode presents a vast array of character choices for the digital humanist, so many that it can be difficult to figure out which character – if any – is the appropriate one to use. When working on a digital version of a Latin text that contains Roman numerals, should text encoder use U+2160 ROMAN NUMERAL ONE or U+0049 LATIN CAPITAL LETTER I? Should one use the duplicate ASCII characters that are located at U+FF01ff. (and why were they ever approved)? These types of questions can create confusion for text encoders.

The process of approving new characters by the Unicode Technical Committee and the relevant ISO committee is intended to be open, meaning that scholars, representatives of companies and national bodies, and other individuals may make proposals and, to a certain extent, participate in the process. Yet which characters get approved – and which don't – can still be baffling. On the TEI-list, one member wrote on 1 August 2009: "What is and isn't represented in unicode is largely a haphazard mishmash of bias, accident and brute-force normalisation. Unicode would be dreadful, if it weren't for the fact that all the alternatives are much worse."

This paper addresses the question of which characters get approved and which don't, by examining the forces at work behind the scenes, issues about which digital humanists may not be aware. This talk, by a member of the two standards

committees on coded character sets, is meant to give greater insight into character encoding today, so that the character encoding standard doesn't seem like a confusing set of decisions handed down from a faceless group of non-scholars. Specific examples of the issues will be given and the talk will end with suggestions so that digital humanists, armed with such information, will feel more confident in putting forward proposals for needed, eligible characters.

The Unicode Technical Committee (UTC) is one of the two standards committees that must approve all new characters. Since it is composed primarily of industry representatives, technical discussion often predominates at meetings, including the question of whether given characters (or scripts) can be supported in current font technology and in software. For the academic, the question of whether a given character (or script) can be implemented in current fonts/software is not one commonly considered, and wouldn't necessarily be known, unless they attended the UTC meetings in person. Also, the acceptance of characters can be based on current Unicode policy or precedence of earlier encoding decisions, which again is often not known to outsiders. How to handle historical ligatures, for example, has been discussed and debated within UTC meetings, but since the public minutes of the UTC do not include the discussion surrounding a character proposal, it may appear that the UTC is blind to scholars' concerns, which is frequently not the case. In order to have a good chance at getting a proposal approved in the UTC, it is hence important for scholars to work with a current member of the UTC who can trouble-shoot proposals and act as an advocate in the meetings, as well as explain concerns of the committee and the reasoning behind their decisions.

The ISO/IEC JTC1/SC2 Working Group 2, a working group on coded character sets, is the second group that must approve characters. This group is composed of national standards body representatives. Unlike the UTC, the WG2 is not primarily technical in nature, as it is a forum where national standards bodies can weigh in on character encoding decisions. This group is more of a "United Nations" of character encoding, with politics playing a role. Discussion can, for example, involve the names of characters and scripts, which can vary by language and country, thus causing disagreement among member bodies. Like the other International Organization for Standardization groups, decisions are primarily done by consensus (International Organization for Standardization, "My ISO Job: Guidance for delegates and experts", 2009). This means that within WG2, disagreements amongst members can stall a particular character or script proposal from being approved. For example, a proposal for an early script used in Hungary

is currently held up in WG2, primarily because there is disagreement between the representatives from Austria (and Ireland) and the representative from Hungary over the name. To the scholar, accommodating national standards body positions when making encoding decisions may seem like unnecessary interference from the political realm. Still, diplomatic concerns need to be taken into account in order for consensus to be reached so proposals can be approved. Again, having the support of one's national body is a key to successfully getting a character proposal approved.

Since WG2 is a volunteer standards organization within ISO, it relies on its members to review proposals carefully, and submit feedback. Unfortunately, many scholars don't participate in ISO, partly because it involves understanding the international standard development process, as well as a long-term commitment – the entire approval process can take at least two years. Another factor that may explain the lack of regular academic involvement is that scholars participating in standards work typically do not receive professional credit. Because there is not much expert involvement in WG2 to review documents (perhaps even fewer experts than in the UTC), errors can creep in. For many of the big historic East Asian script proposals, for example, only a small handful of people are reviewing the documents, which is worrisome. The recently addition of CJK characters ("Extension C"), which has 4,149 characters, could have benefited from more scholarly review. Clearly there remains a critical need for specialists to become involved in the ISO Working Group 2, so as to prevent the inclusion of errors in future versions of the character encoding standard.

Besides the activity within each separate standards group, there are developments affecting both standards groups that may not be known to digital humanists, but which influence character encoding. New rules have recently been proposed within ISO, for example, which will slow the pace at which new characters and scripts are approved by ISO and published in *The Unicode Standard* (ISO/IEC JTC1/SC2 meeting, 2009). The new rules will severely impact new character requests. Another example of activity affecting digital projects, particularly those using East Asian characters, was the announcement in October 2009 by the Japanese National Body that it has withdrawn its request for 2,621 rare ideographs ("gaiji" characters) (Japan [National Body], "Follow-up on N3530 (Compatibility Ideographs for Government Use)", 2009), instead opting to register them in the Ideographic Variation Database, a Unicode Consortium-hosted registry of variation sequences that contain unified ideographs (Unicode Consortium, "Ideographic Variation Database",

2009). The use of variation selectors is a different approach than that advocated in the TEI P5 for "gaiji" characters (TEI P5 Guidelines: "5. Representation of Non-standard Characters and Glyphs"), but is one that should be mentioned in future *TEI Guidelines* as an alternative.¹ In order to keep apprised of developments within the standards groups, a liaison between TEI and the Unicode Consortium (and/or ISO/IEC JTC1/SC2) would be advisable, as the activities of Unicode (/ISO) can influence TEI recommendations.

In sum, the process of character encoding is one that ultimately involves people making decisions. Being aware of the interests and backgrounds of each standard group and their members can help explain what appears to be a spotty set of characters in Unicode. Keeping up-to-date on developments within the committees can also provide insight into why a particular character is approved or not, or why its progression has been slowed. The talk will conclude with suggestions on how digital humanists can participate more actively and effectively in the standards process.

References

- Davis, Mark** (May 2009). *Moving to Unicode 5.1*. <http://googleblog.blogspot.com/2008/05/moving-to-unicode-51.html> (accessed 15 November 2009).
- Unicode Consortium.** *Unicode 5.2.0*. 21 October 2009 http://www.unicode.org/versions/Unicode_5.2.0/ (accessed 15 November 2009).
- International Organization for Standardization.** *My ISO Job: Guidance for delegates and experts*. http://www.iso.org/iso/my_iso_job.pdf (accessed 15 November 2009).
- ISO/IEC JTC1/SC2 meeting*. Tokyo, Japan, 30 October 2009.
- Japan [National Body]** (16 October 2009). *Follow-up on N3530 (Compatibility Ideographs for Government Use)*. <http://std.dkuug.dk/jtc1/sc2/wg2/docs/N3706.doc>.
- Unicode Consortium.** *Ideographic Variation Database*. <http://www.unicode.org/ivd/> (accessed 15 November 2009).
- TEI Consortium (ed.)**. 'P5. Representation of Non-standard Characters and Glyphs'. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.5.0*. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/WD.html> (accessed 15 November 2009).

Notes

1. Variation selectors have also been mentioned as being used to handle variants in other scripts, such as the historic script Tangut.

Semantic Cartography: Using RDF/OWL to Build Adaptable Tools for Text Exploration

Ashton, Andrew

Andrew_Ashton@brown.edu

Center for Digital Scholarship, Brown University,
USA

Texts encoded using the Text Encoding Initiative Guidelines (TEI) are ideally suited to close examination using a variety of digital methodologies and tools. However, because the TEI is a broad set of guidelines rather than a single schema, and because encoding practices and standards vary widely between collections, programmatic interchange among projects can be difficult. Software that operates effectively across TEI collections requires a mechanism for normalizing data - a mechanism that, ideally, preserves the essence of the source encoding. Brown University's Center for Digital Scholarship will address this issue as one part of a project, funded by a National Endowment for the Humanities Digital Humanities Start-Up Grant, to develop tools for analyzing TEI-encoded texts using the SEASR environment (National Center for Supercomputing Applications).

SEASR is a scholarly software framework that allows developers to create small software modules that perform discrete, often quite simple, analytical processing of data. These modules can be chained together and rearranged to create complex and nuanced analyses. Both the individual modules (or components, in SEASR's parlance) and the chains (or flows) running within a SEASR server environment can be made available as web services, providing for a seamless link between text collections and the many visualizations, analysis tools, and APIs available on the web. For literary scholars using TEI, this approach offers innumerable possibilities for harnessing the semantic richness encoded in the digital text, provided that there is a technological mechanism for negotiating the variety of encoding standards and practices typical of TEI-based projects.

The central thrust of Brown's effort is to develop a set of SEASR components that exploit the semantic detail available in TEI-encoded texts. These components will allow users to submit texts to a SEASR service and get back a result derived from the analytical flow that they have constructed. Results could include a data set describing morphosyntactic

features of a text, a visualization of personal relationships or geographic references, a simple breakdown of textual features as they change over time within a specific genre, etc. SEASR flows can also be used to transform parts of TEI documents into more generic formats in order to use data from digital texts with web APIs, such as Google Maps. Because these components deal with semantic concepts rather than raw, predictable data, they require a mechanism to map concepts to their representations in the encoded texts. Furthermore, in order for these modules to be applicable to a variety of research collections, they must be able to adapt to different manifestations of semantically similar information. Users need the freedom to tell the software about the ways in which data with semantic meaning, such as relationships or sentiment, are encoded in their texts. To do this, we will build a semantic map – a document that describes a number of relationships between 1) software (in this case, SEASR components), 2) ontologies, and 3) encoded texts (see Figure 1).

Example 1.

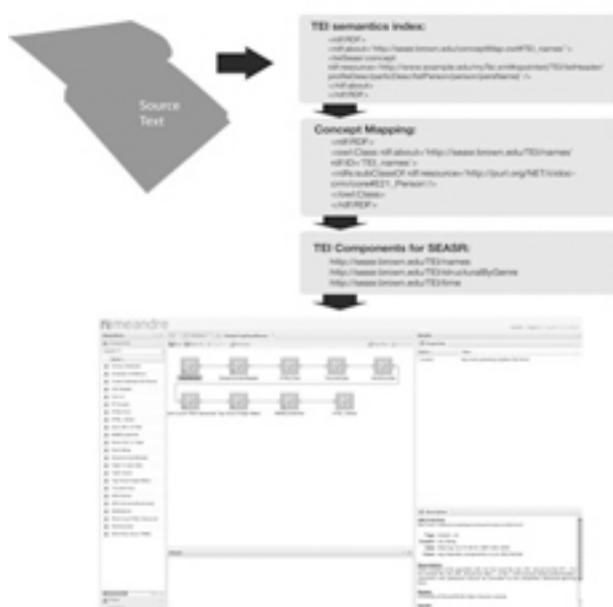


Figure 1

The semantic map is modeled on the approaches of previous projects working with semantically rich TEI collections, most notably the work of the Henry III Fine Rolls Project, based at King's College London. The creators of that project developed a method for using RDF/OWL to model the complex interpersonal and spatial relationships represented in their TEI documents (Vieira and Ciula). They use semantic-web ontologies, such as CIDOC-CRM and SKOS, to define relationships between TEI nodes. Vieira and Ciula offer examples in which TEI nodes that describe individuals are related to other nodes that describe professions, using the CIDOC-CRM ontology to codify the relationship. (Vieira and Ciula, 5) This

approach serves well as a model for mapping some of the more nebulous concepts of interest to the SEASR components to the variable encoding of those concepts in TEI.

The links between the encoded texts and the SEASR components are forged using two techniques that exploit the Linked Data concepts around which SEASR - and many other emerging software tools - are designed (Berners-Lee). The first is a RDF/OWL dictionary that defines the ontology of the semantic concepts at work in the TEI component suite. RDF/OWL allows us to make assertions about the concepts associated with any resource identified by a URI (Uniform Resource Identifier). Within SEASR, any component or flow is addressable via a URI, making it the potential subject of a RDF/OWL expression. For example, a component that extracts personal name references from a text can be defined in an RDF/OWL expression as having an association with the concept of a personal name, as defined by any number of ontologies. The second part of the semantic map is an editable configuration document that uses the XPointer syntax to identify the fragments of a particular TEI collection that correspond to the semantic definitions expressed in the RDF/OWL dictionary. In this example, collections that encode names in several different ways can specify, via XPointer, the expected locations of relevant data. When the SEASR server begins an analysis, data is retrieved from the TEI collection using the parameters defined in the semantic map, and is then passed along to other components for further analysis and output.

In addition to the semantic map and the set of related analytical components, the planned TEI suite for SEASR includes components to ease the retrieval and validation of locally defined data as they are pulled into analytical flows. One such component examines which of the TEI-specific components in a flow have definitions in the RDF/OWL dictionary. The result is passed to another component, which uses Schematron to verify that the locations and relationships expressed in the semantic map are indeed present in the data being received for analysis. A successful response signals SEASR to proceed with the analysis, while an unsuccessful one returns information to the user about which parts of the text failed to validate.

Our approach is different from that of other projects, such as the MONK Project, which have also wrestled with the inevitable variability in encoded collections (MONK Project). For MONK, this issue was especially prominent as the goal of the project was to build tools that could combine data from diverse collections and analyze them as if they were a uniform corpus. This meant handling not only TEI of various flavors, but other types of XML and

SGML documents as well. To solve this problem, MONK investigators developed TEI Analytics, a generalized subset of TEI P5 features designed "to exploit common denominators in these texts while at the same time adding new markup for data structures useful in common analytical tasks" (Pytlak-Zillig, 2009). TEI-A enables developers to combine different text collections to allow large-scale analysis by systems such as MONK. As a solution to handling centralized, large-scale data analysis, TEI-A is an invaluable achievement in light of the maturation of mass-digitization efforts such as Google Books and HathiTrust. However, our goal in creating SEASR components is fundamentally different than that of MONK, and thus warrants a different approach. Our SEASR tools are designed to be shared among institutions but to be used differently by each, as a part of a web interface for a particular collection. Furthermore, the granularity of the concepts of interest to the TEI components for SEASR makes it infeasible that we could easily map such encoding to a common format, such as TEI-A. Hence, the semantic map – an index that acts as a small-scale interpreter between local collections and the abstract semantic notions marked-up within them.

In developing TEI tools for SEASR, we address several issues of immediate interest to scholars and tools-developers in the Digital Humanities. It is certainly useful to create text analysis tools for this new software environment. But of broader interest to scholars and users of digital text collections are the semantic mechanisms that permit interplay between community-based tools and their own collections. Several issues require close scrutiny as this model develops: with careful forethought, we need to ensure that our tools are viable outside of the SEASR framework; we need to consider whether the XPointer syntax has a future in the evolving Semantic Web ecology, and likewise consider how our curated scholarly collections can interact more seamlessly with that environment. Ultimately, the tools that we develop will be available in a public repository for institutions experimenting with SEASR.

Funding: This work was supported by the National Endowment for the Humanities Digital Humanities Start Up Grants program.

National Center for Supercomputing Applications (NCSA). SEASR: Software Environment for the Advancement of Scholarly Research. <http://www.s easr.org> (accessed 14 November, 2009).

Pytlak-Zillig, B. (2009). 'TEI Analytics: converting documents into a TEI format for cross-collection text analysis'. *Literary and Linguistic Computing*. **24(2)**: 187-192. <http://doi:10.1093/linc/fqp005>.

Vieira, J.M. and Ciula, A. (2007). 'Implementing an RDF/OWL Ontology on Henry the III Fine Rolls'. *OWLED 2007*. Innsbruck, Austria, June 2007. http://www.webont.org/owled/2007/Pape rsPDF/submission_6.pdf.

References

- Berners-Lee, T..** *Linked Data – Design Issues*. <http://www.w3.org/DesignIssues/Linke dData.html> (accessed 14 November, 2009).
- MONK Project.** <http://www.monkproject.org> (accessed 14 November, 2009).

Using Wikipedia to Enable Entity Retrieval and Visualization Concerning the Intellectual/Cultural Heritage

Athenikos, Sofia J.

sofia.j.athenikos@acm.org

Drexel University, Philadelphia, USA

At the 2009 Digital Humanities conference I presented my paper on the WikiPhiloSofia (<http://research.cis.drexel.edu:8080/sofia/WPS/>) project (Athenikos and Lin, 2009), which was concerned with extraction and visualization of facts, relations, and networks concerning philosophers using Wikipedia (<http://www.wikipedia.org/>) as the data source. In this proposal, I present a related, extended project in progress, entitled PanAnthropon, which incorporates the problems of retrieving entities in response to a query and retrieving entities related to a given entity and which extends the scope of application to domains other than philosophy.

1. Background

Traditional information retrieval is concerned with retrieving documents that are potentially relevant to a user's query. The relevance of a document to a given query is usually measured by lexico-syntactic matching between the terms in the query and those in the document (title). Familiar Web search engines, such as Google and Yahoo, for example, return a ranked list of Web pages that contain all or some of the keywords in the query entered by a user. The Semantic Web (Berners-Lee et al., 2001) initiative aims at transforming the Web of pages (documents) into the Web of entities (things in the broadest sense) (cf. OKKAM project (<http://www.okkam.org/>) (Bouquet et al., 2007)). Information retrieval on the Semantic Web is no longer a matter of retrieving documents via semantics-unaware keyword matching but a matter of retrieving entities that satisfy the semantic constraints imposed by the query, i.e. those that are of specific semantic type and that satisfy the given semantic conditions. Wikipedia has become an important semantic knowledge resource (cf. Zesch et al., 2007) thanks to its unique set of semi-structured semantic features and the huge amount of content covering a wide range of topics. What renders Wikipedia more interesting is the fact that it

can be considered as a self-contained web of entities. Each Wikipedia article is concerned with one entity, and the given entity is connected to other entities via explicit semantic relations as in infoboxes and wikitables or via implicit semantic relations as in hyperlinks.

2. Motivation

Through the WikiPhiloSofia project I demonstrated extracting, retrieving, and visualizing specific facets of information, not documents, concerning entities of a selected type, namely, philosophers, by exploiting the hyperlinks, categories, infoboxes, and wikitables contained in Wikipedia articles. The interface that I created enables the users to select a focus of query in the form of an entity (philosopher) or a pair of entities (philosophers) and then to retrieve entities that satisfy specified conditions with respect to the given entity or pair of entities. However, the project did not consider the problems of retrieving entities as answers to queries, semantically typing entities, or retrieving related entities by type and condition.

The proposed PanAnthropon project takes up the aforementioned problems left out of the WikiPhiloSofia project. The dual objective is to enable retrieval of entities that directly answer a given semantics-based query and to enable retrieval of related entities by semantic type, subtype (role), and relation, by using information extracted/integrated from Wikipedia. The project applies the approach to the entities concerning the intellectual/cultural heritage – people, works, concepts, etc. The Web portal interface thereby constructed will allow the users to retrieve entities that directly answer their queries as well as to explore people, works, concepts, etc. *in relation to* other people, works, concepts, etc.

3. Conceptualization

In the proposed project, "entities" are conceived of as things of all kinds that have certain properties (or attributes). The "type" of an entity is considered as a generic kind (or class or category) into which the given entity is classified, e.g., person, work, etc. In general, the type of an entity is fixed and exclusive in the sense that an entity that belongs to one type does not or cannot belong to other types. The "subtype" of an entity refers to a subclass or subcategory into which the entity can be classified, under a given type. The subtype of an entity is fluid and non-exclusive in the sense that an entity may belong to more than one subtype (under a given type). This is especially so in the case of person-type entities, and thus a subtype may better be understood as a "role" in this case. In general, there are multiple subtypes under a given type, and the former can be further classified into still more specific subtypes. A type or subtype of an entity

can be considered as a special kind of property. A “fact” concerning an entity refers to a tuple consisting of <entity, attribute, value, [context]>, which adds the optional “context” element to the <subject, predicate, object> triple model. An entity can have “relations” to other entities, given its properties. The kinds of properties and relations that are relevant or of interest concerning an entity, except certain basic facts, depend on the domain at issue. An entity may belong to multiple domains, but not every subtype, property, or relation is relevant or equally important in one domain as in another domain. The project therefore intends to build a portal consisting of sub-portals representing different domains.

4. Methodology

4.1. Data Extraction and Processing

The pre-processing stage of the project (for each domain) concerns: (1) compiling a seed list of entities of interest by extracting names from various lists and categories in Wikipedia, (2) downloading Wikipedia article pages for each entry on the list, and (3) inspecting typical attributes and (types/subtypes of) values contained in infoboxes, wikitables, etc. The main processing stage concerns extracting information on an entity and related entities from each Wikipedia page. The semantic type/subtypes of a given entity are extracted and/or assigned. Semi-structured templates and portions of the article are processed so as to extract attribute–value (or predicate–object or relation–entity) pairs. The related entities are matched to the entities on the seed entity list. Additional Wikipedia pages are downloaded for entities not matched on the list, and the information on those entities is extracted. The (optional) post-processing stage concerns converting the data stored in a MySQL database to XML files and RDF triples, thereby creating a semantic data repository that can be linked to other resources involved in the Linked Data (<http://linkeddata.org/>) initiative in the latter case.

4.2. Semantic Search Interface

The semantic search interface created will support three types of query and retrieval. The first type of query/retrieval concerns retrieval of entities that correspond to queries the expected answers of which are entities. The second type of query/retrieval concerns retrieval of entities related to an entity, according to type/subtypes and specified properties/values. The third type of query/retrieval concerns retrieval of facts concerning an entity. The interface will also incorporate some of the visualization features available in the WikiPhiloSofia portal interface.

5. Current Application

The film domain has been chosen as the initial proof-of-concept domain of application. In my presentation I will demonstrate the entity retrieval functionalities with 1.5+ million (and growing) facts about 11370 films, 69545 persons, 74545 film roles, 253 places, 6033 dates, etc.

6. Related Work

The task of retrieving entities in response to user queries using the information in Wikipedia has since 2007 been the focus of the INEX (Initiative for the Evaluation of XML Retrieval) XML Entity Ranking (XER) Track (de Vries et al., 2008; Demartini et al., 2009). Unlike in the INEX XER Track, the proposed project addresses the task by extracting information from the HTML Wikipedia files and building a knowledge base based on it. The task of constructing a knowledge base by extracting information from templates in Wikipedia such as infoboxes has been attempted in large scale by, e.g., Auer and Lehmann (2007) and Suchanek et al. (2007). There is also the DBpedia (<http://dbpedia.org/>) knowledge base, which contains the information extracted from Wikipedia. The proposed project, however, utilizes the information in the main content of Wikipedia articles, as well as templates, to enable and enhance entity retrieval. It will also provide a more flexible working search interface for both general and entity-specific queries.

7. Conclusion

The PanAnthropon project utilizes Wikipedia as a semantic knowledge source for entity retrieval and applies the approach to materials concerning the intellectual/cultural heritage. The semantic search interface created will allow the users to retrieve entities that directly answer their queries as well as to explore various semantic facets concerning those entities. As such, it will provide a useful resource for digital humanities.

References

- Athenikos, S.J., Lin, X.** (2009). 'WikiPhiloSofia: Extraction and Visualization of Facts, Relations, and Networks Concerning Philosophers Using Wikipedia'. *Conference Abstracts of Digital Humanities 2009*. Pp. 56-62.
- Auer, S., Lehmann, J.** (2007). 'What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content'. *Proceedings of 4th*

European Semantic Web Conference (ESWC 2007).
Innsbruck, Austria, June 2007.

Berners-Lee, T., Hendler, J., Lassila, O.
(2001). 'The Semantic Web'. *Scientific American*. 5
(May 2001).

Bouquet, P., Stoermer, H., Giacomuzzi, D.
(2007). 'OKKAM: Enabling a Web of Entities'.
*Proceedings of the 16th International World Wide
Web Conference (WWW 2007)*. Banff, Alberta,
Canada, 8-12 May 2007.

**Demartini, G., de Vries, A.P., Iofciu, T., Zhu,
J.** (2009). *Overview of the INEX 2008 Entity
Ranking Track, INEX 2008*. LNCS. Heidelberg:
Springer-Verlag, Berlin. V. 5631, pp. 243-252.

**de Vries, A.P., Vercoustre, A.-M., Thom, J.A.,
Craswell, N., Lalmas, M.** (2008). *Overview of
the INEX 2007 Entity Ranking Track, INEX 2007*.
LNCS. Heidelberg: Springer-Verlag, Berlin. V. 4862,
pp. 245-251.

Suchanek, F.M., Kasneci, G., Weikum, G.
(2007). 'YAGO: A Core of Semantic Knowledge
Unifying WordNet and Wikipedia'. *Proceedings of
the 16th International World Wide Web Conference
(WWW 2007)*. Banff, Alberta, Canada, pp. 697-706.

Zesch, T., Gurevych, I., Mühlhäuser, M.
(2007). 'Analyzing and Accessing Wikipedia
as a Lexical Semantic Resource'. *Proceedings
of the Biannual Conference of the Society
for Computational Linguistics and Language
Technology*. Tübingen, Germany, April 2007.

Mapping the World of an Ancient Greek Historian: The HESTIA Project

Barker, Elton

e.t.e.barker@open.ac.uk
Open University, UK

Pelling, Chris

chris.pelling@chch.ox.ac.uk
University of Oxford, UK

Bouzarovski, Stefan

BuzarS@adf.bham.ac.uk
University of Birmingham, UK

Isaksen, Leif

l.isaksen@soton.ac.uk
University of Southampton, UK

HESTIA (the Herodotus Encoded Space-Text-Imaging Archive) is an interdisciplinary project, sponsored by the AHRC and involving the collaboration of academics from Classics, Geography and Archaeological Computing, that aims to enrich contemporary discussions of space by developing an innovative methodology to the study of an ancient narrative, Herodotus' *Histories*. Using the latest ICT, it investigates the ways in which space is represented in the *Histories*, and develops visual tools to capture the 'deep' topological structures of the text, extending beyond the usual two-dimensional Cartesian maps of the ancient world. In addition to exploring the network culture that Herodotus represents, one of its stated outcomes is to introduce Herodotus' world to new audiences via the internet. This paper will set out in more detail that methodology, paying particular attention to the decisions that we have made and the problems that we have encountered, in the hope that our project can contribute not only to offering a more complex picture of space in Herodotus but also to establishing a basis for future digital projects across the humanities which deal with large text-based corpora.

For the purposes of a twenty minute presentation, we address three key areas of interest:

1. To provide the background to the data capture and digital mark-up of the *Histories*. Our project differs from many by utilizing a digital resource already in the public domain: the text of Herodotus freely available from Perseus (<http://www.perseus.tufts.edu/hopper/>). Though the capture of the digital text from Perseus (version P4) gave our project a welcome initial boost, a number of issues

have had to be overcome including procedural conversion, which involves handing back a P5 text to Perseus.

2. To sketch out the type, structure and categorization of our spatial database. A PostgreSQL database was chosen because its PostGIS extension provides excellent functionality for spatial data and is widely supported by other applications: one key principle of HESTIA has been to use open source software in order to maximize its potential dissemination and reusability of its data. By storing information about references, locations and the text in the database, it has been possible to provide it to both a Desktop GIS system and Webmapping server simultaneously (see figure 1).
3. To present a sample set of results of the maps that we have been able to generate using the geo-referenced database. While sections 1 and 2 will be of particular concern to anyone wishing to understand how one may interrogate spatial data using the digital resources available, this last stage holds the greatest interest for the non ICT expert since it demonstrates the use to which data in this form can be employed: hence the main focus of this paper will be on explaining the five kinds of map that we have been able to generate:
 - i. *Geographical Information System (GIS) maps.* The most basic maps that are generated simply represent a ‘flat’ image of the spatial data: that is to say, they mark all the places that Herodotus mentions over the course of his work with a single point, thereby providing a snapshot of the huge scope of his enquiry. In this way one is able to gain an overview of the places mentioned in Herodotus and divide them according to three different kinds of spatial category: settlement, territory and physical feature (see figure 2). A variation on this basic model depicts places according to the number of times they are mentioned (see figure 3).
 - ii. *GoogleEarth.* In order to start experimenting with public dissemination it was decided to expose the PostGIS data as KML: a markup format that can be read by a variety of mapping applications including GoogleEarth. With this ‘Herodotus Earth’ application, users will be able to construct ‘mashups’ of visual and textual data. So, for example, since all places are linked to entries in the database, when one clicks on a particular location in GoogleEarth, it will be possible to bring up a dialog box containing Herodotus’ text (in both English and Greek) for that particular location

for every occasion when it is mentioned in the narrative (see figure 4).

- iii. *TimeMap.* Whilst it is possible to visualise narrative change using graphs, and static differences using GIS, it is more difficult to visualize spatial changes throughout the narrative; GIS does not have useful functionality in this regard except for the ability to turn layers on and off, a process which becomes impractical beyond book level. The most likely candidate to provide this kind of functionality is an Open Source JavaScript project called TimeMap, developed by Nick Rabinowitz, which draws on several other technologies in order to allow data plotted on GoogleMaps to appear and disappear as a timeline is moved. In collaboration with the project’s IT consultant Leif Isaksen, Nick Rabinowitz has adapted his schema in order to represent the book structure of Herodotus’ narrative in a similar way (see figure 5).
- iv. *Database-generated network maps.* Since the GIS maps outlined in i. have little to say per se regarding Herodotus’ organization of space, a key next step has been to explore rapidly-generated networks based on the simple co-presence of terms within sections of the text. The purpose of producing networks of this kind is to start exploring the connections that Herodotus himself makes between places, seeing how strongly the narrative is bound to geographical regions, and flagging up potential links between particular locations (see figure 6). Figure 7 illustrates one such simple network, that for “territories” across the entire *Histories*. It shows a series of links connecting Greece to other areas within the Mediterranean world: but the territory that has the strongest connections in this basic network culture is Egypt. While surprising, it does make sense on reflection, since for a better part of one book Herodotus uses Egypt as the touchstone against which other cultures, including Persia and his own, Greece, are compared. It is as a tool of comparison, then, that Egypt appears to be the centre of Herodotus’ network picture of the Mediterranean. Figure 8 complements this picture by presenting the networks of physical features, which envelop the comparison between Greece and Egypt.
- v. *Manual network maps.* The automated maps outlined in iv. rely on ‘counting’ the number of times two or more places are connected to each other: they have little to say about the kind of connection being drawn. We end our presentation, then, with a brief comparison

to text-based qualitative analysis, which attempts to categorize relationships according to fundamental geographical concepts of movement or transformation, based on the close reading of one book (5). Our different approaches are intended to complement, challenge and inform each other with a view also to suggest ways by which the automated process may be extended, such as by adopting text-mining procedures.

In sum, this paper aims to meet three outcomes:

1. To outline a methodology for dealing with digital data that may be transferred, adapted and improved upon in other fields of digital humanities.
2. To demonstrate the value of digital projects within the humanities for helping to achieve ‘impact’ by bringing the world of a fifth-century BC Greek historian into everyone’s home.
3. To show the potential for the digital manipulation of data in posing new kinds of research questions.



A screenshot of a digital interface, possibly a database or a specialized software. It displays a grid of data with columns labeled 'Name', 'Age', 'Gender', 'Occupation', 'Location', and 'Notes'. The data includes entries for various individuals, such as 'Athena' (Age 20, Gender Female, Occupation Goddess, Location Athens, Note: Queen of Gods), 'Zeus' (Age 100, Gender Male, Occupation King of Gods, Location Olympia, Note: Father of Gods), and 'Hercules' (Age 30, Gender Male, Occupation Hero, Location Corinth, Note: Son of Zeus).

Fig. 1

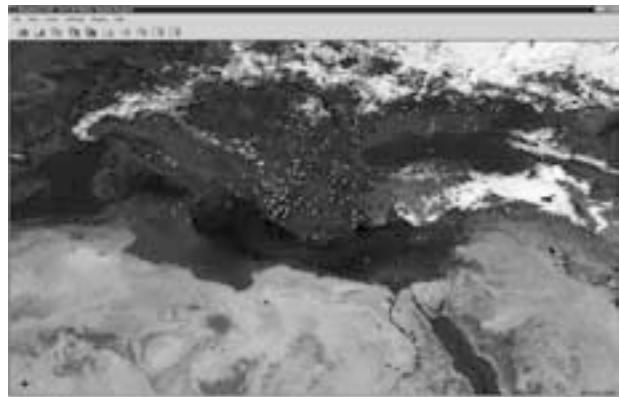


Fig. 2

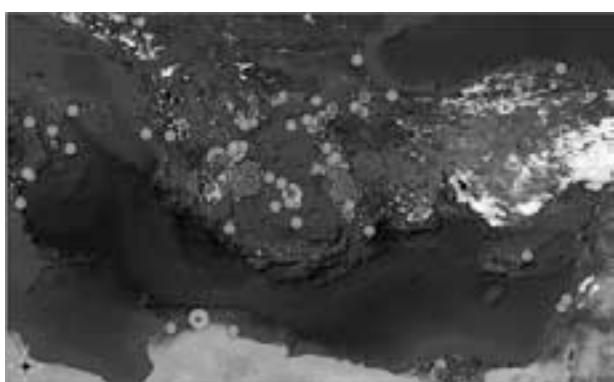


Fig. 3



Fig. 4



Fig. 5

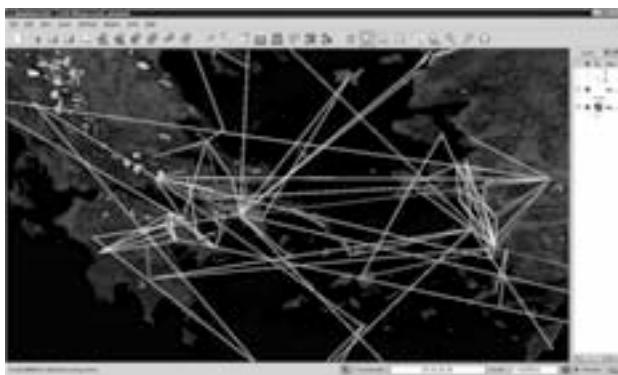


Fig. 6

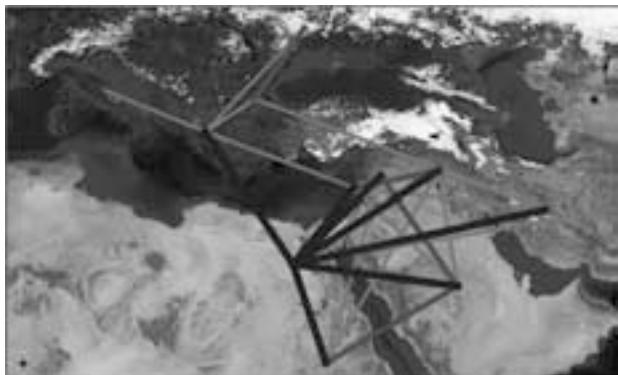


Fig. 7

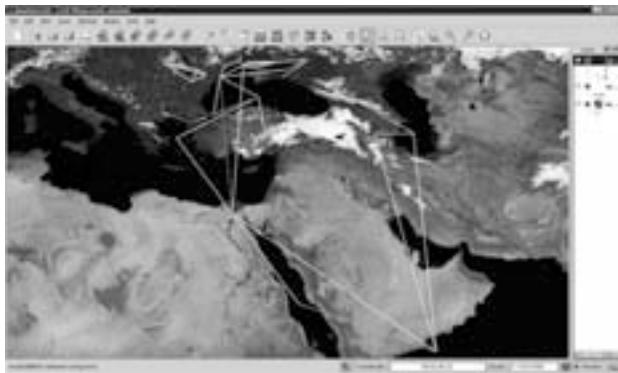


Fig. 8

Horden, J., Purcell, N. (2000). *The Corrupting Sea*. Oxford.

Jackson, P. (1994). *Maps of Meaning: An Introduction to Cultural Geography*. London.

Janni, P. (1984). *La Mappa e il Periplo. Cartografia antica e spazio odologico*. Marcerata.

Kwan, M.-P., Ding, G. (2008). 'Geo-Narrative: Extending Geographic Information Systems for Narrative Analysis in Qualitative and Mixed-Method Research'. *The Professional Geographer*. **60.4**: 443-65.

Lefebvre, H. (1991). *The Production of Space*. Chicago.

Lloyd, C. D., Lilley, K. D. (2009). 'Cartographic Veracity in Medieval Mapping: Analyzing Geographical Variation in the Gough Map of Great Britain'. *Annals of the Association of American Geographers*. **99.1**: 27-48.

Purves, A. C. (2002). *Telling Space: Topography, Time and Narrative from Homer to Xenophon*. University of Pennsylvania Ph.D. Dissertation.

Romm, J.S. (1994). *The Edges of the Earth in Ancient Thought: Geography, Exploration, and Fiction*. Princeton.

Sheppard, E. (2005). 'Knowledge Production through Critical GIS: Genealogy and Prospects'. *Cartographica: The International Journal for Geographic Information and Geovisualization*. **40.4**: 5-21.

Tuan, Y.-F. (1978). 'Literature and geography: implications for geographical research'. *Humanistic Geography*. Ley, Samuels (eds.). London, pp. 194-206.

References

Berman, M. L. (2005). 'Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History'. *Historical Geography*. **33**: 118-33.

Crampton, J.W., Krygier, J. (2006). 'An introduction to critical cartography'. *ACME: An International E-Journal for Critical Geographies*. **4**: 11-33.

Harley, J.B. (1989). 'Deconstructing the map'. *Cartographica: The International Journal for Geographic Information and Geovisualization*. **26.2**: 1-20.

TEI P5 as a Text Encoding Standard for Multilevel Corpus Annotation

Bański, Piotr

pkbanski@uw.edu.pl

University of Warsaw

Adam Przepiórkowski

adamp@ipipan.waw.pl

Institute of Computer Science Polish Academy of Sciences

The need for text encoding standards for language resources (LRs) is widely acknowledged: within the International Standards Organization (ISO) Technical Committee 37 / Subcommittee 4 (TC 37 / SC 4), work in this area has been going on since the early 2000s, and working groups devoted to this issue have been set up in two current pan-European projects, CLARIN (<http://www.clarin.eu/>) and FLaReNet (<http://www.flarenet.eu/>). It is obvious that standards are necessary for the interoperability of tools and for the facilitation of data exchange between projects, but they are also needed within projects, especially where multiple partners and multiple levels of linguistic data are involved.

One such project is the National Corpus of Polish (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; <http://nkjp.pl/>; Przepiórkowski *et al.* 2008, 2009) involving 4 Polish institutions and carried out in 2008–2010. The project aims at the creation of a 1-billion-word automatically annotated corpus of Polish, with a 1-million-word subcorpus annotated manually. The following levels of linguistic annotation are distinguished in the project: 1) segmentation into sentences, 2) segmentation into fine-grained word-level tokens, 3) morphosyntactic analysis, 4) coarse-grained syntactic words (e.g., analytical forms, constructions involving bound words, etc.), 5) named entities, 6) syntactic groups, 7) word senses (for a limited number of ambiguous lexemes).

Any standards adopted for these levels should allow for stand-off annotation, as is now common practice and as is virtually indispensable in the case of many levels of annotation, possibly involving conflicting hierarchies.

Two additional, non-linguistic levels of annotation required for each document are text structure (e.g., division into chapters, sections and paragraphs, appropriate marking of front matter, etc.) and metadata. The standard adopted for these levels

should be sufficiently flexible to allow for representing diverse types of texts, including books, articles, blogs and transcripts of spoken data.

NKJP is committed to following current standards and best practices in corpus development and text encoding. However, because of the current proliferation of official, *de facto* and purported standards, it is far from clear what standards a new corpus project should adopt. The aim of this paper is to attempt to answer this question.

1. Standards and best practices

The three text encoding standards and best practices listed in a recent CLARIN short guide (CLARIN:STE, 2009)¹ are: standards developed within ISO TC 37 SC 4, the Text Encoding Initiative (TEI; Burnard and Bauman 2008) guidelines and the XML version of the Corpus Encoding Standard (XCES; Ide *et al.* 2000). Apart from these, there are other *de facto* standards and best practices, e.g., TIGER-XML (Mengel and Lezius, 2000) for the encoding of syntactic information, or the more general PAULA (Dipper, 2005) encoding schema used in various projects in Germany.

1.1. XCES

The original version of XCES inherits from TEI an exhaustive approach to metadata representation. It makes specific recommendations for the representation of morphosyntactic information and for the alignment of parallel corpora. In early the 2000s, it was probably the most popular corpus encoding standard.

Currently, the claim of XCES to being such a standard is much weaker. A new – more abstract – version of XCES was introduced around 2003, where concrete morphosyntactic schema was replaced by a general feature structure mechanism, different from the ISO Feature Structure Representation (FSR) standard (ISO 24610-1). In our view, this is a step back, as adopting a more abstract representation requires more work on the part of corpus developers. Moreover, XCES has no specific recommendations for other levels of linguistic knowledge, and no mechanisms for representing discontinuity and alternatives, all of which need to be represented in NKJP. Taking also into account the lack of documentation and the potential confusion concerning its versioning,² XCES turns out to be unsuitable for the purposes of NKJP.

1.2. ISO TC37 SC 4

There is a family of ISO standards developed by ISO TC 37 SC 4 for modelling and representing

different types of linguistic information. The two published standards concern the representation of feature structures (ISO 24610-1) and the encoding of dictionaries (ISO 24613). Other proposed standards are at varying levels of maturity and abstractness. While eventually these standards may reach stability and specificity required by practical applications, this is currently not the case.³

1.3. TIGER-XML and PAULA

TIGER-XML and a schema which may be considered as its generalisation, PAULA, are specific, relatively well-documented and widely employed best practices for describing linguistic objects occurring in texts (so-called "markables") and relations between them (in the case of TIGER-XML, the constituency relation). They do not contain specifications for metadata or structural annotation.

2. TEI P5

For metadata and structural annotation levels there is no real alternative to TEI. Moreover, TEI P5 implements the FSR standard ISO 24610-1, which can be used for the representation of any linguistic content, along the lines of XCES (although the feature structure representations used in XCES do not comply with this standard), PAULA and the proposed ISO standard, Linguistic Annotation Framework (ISO 24612). TEI P5 is stable, has rich documentation and an active user base, and for these reasons alone it should be preferred to XCES and (the current versions of) the ISO standards. Moreover, any TIGER-XML and PAULA annotation may be expressed in TEI in an isomorphic way, thanks to the linking mechanisms of TEI P5.

However, TEI is a very rich toolbox, proposing multitudinous mechanisms for representing multifarious aspects of text encoding, and this richness, as well as the sheer size of TEI P5 documentation (1350–1400 pages), are often perceived by corpus developers as prohibitive. For this reason, within NKJP, a specific set of recommendations for particular levels of annotation has been developed, aiming at achieving a maximal compatibility (understood as the easiness to translate between formats) with other proposed and *de facto* standards.

For example, TEI P5 offers, among others, the following ways to represent syntactic constituency:

- XML tree, built with elements such as <s>(entence), <phr>(ase), <cl>(ause) and <w>(ord), may directly mirror constituency tree;
- all information, including constituency, may be encoded as a feature structure (Witt *et al.*, 2009);

- each syntactic group is a <seg>(ment) of type group, containing a feature structure description and <ptr> pointers to other constituents, defined in the same file (for non-terminal syntactic groups) or in a stand-off way elsewhere (for terminal words).

While the first of these representations is the most direct, and the second most general, it is the third representation that directly mirrors TIGER-XML, PAULA and SynAF, and for this reason, it has been adopted in NKJP.

References

- Bel, N., Beskow, J., Boves, L., Budin, G., Calzolari, N., Choukri, K., Hinrichs, E., Krauwer, S., Lemnitzer, L., Piperidis, S., Przepiórkowski, A., Romary, L., Schiel, F., Schmidt, H., Uszkoreit, H., and Wittenburg, P.** (2009). *Standardisation action plan for Clarin. State: Proposal to CLARIN Community*.
- Burnard, L. and Bauman, S. (ed.)** (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. <http://www.tei-c.org/Guidelines/P5/>.
- CLARIN:STE (ed.)** (2009). *Standards for text encoding: A CLARIN shortguide*. <http://www.clarin.eu/documents>.
- Dipper, S.** (2005). 'Stand-off representation and exploitation of multi-level linguistic annotation'. *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, pp. 39-50.
- Ide, N., Bonhomme, P., and Romary, L.** (2000). 'XCES: An XML-based standard for linguistic corpora'. *LREC*. **2000**: 825-830.
- LREC** (2000). 'Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000'. ELRA. Athens.
- Mengel, A. and Lezius, W.** (2000). 'An XML-based encoding format for syntactically annotated corpora'. *LREC*. **2000**: 121-126.
- Przepiórkowski, A., Górska, R. L., Lewandowska-Tomaszczyk, B., and Łaziński, M.** (2008). 'Towards the National Corpus of Polish'. *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*. Marrakech: ELRA.
- Przepiórkowski, A., Górska, R. L., Łaziński, M., and Pezik, P.** (2009). 'Recent developments in the National Corpus of Polish'. *Proceedings of Slovko 2009: Fifth International Conference on NLP, Corpus Linguistics, Corpus Based Grammar*

Research, 25–27 November 2009, Smolenice/Bratislava, Slovakia. Levická, J and Garabík, R. (ed.). Brno. Tribun.

Witt, A., Rehm, G., Hinrichs, E., Lehmburg, T., and Stemann, J. (2009). 'SusTEInability of linguistic resources through feature structures'. *Literary and Linguistic Computing.* 24(3): 363–372.

Notes

1. See also Bel *et al.* 2009.
2. Two different sets of schemata have co-existed on XCES WWW pages since 2003, one given as DTD, another as XML Schema, without any clear indication that they specify different structures.
3. A tendency may be observed of increasing abstractness and generality of proposed standards, esp., SynAF (ISO 24615) and LAF (ISO 24612)". This leads to their greater formal elegance, at the cost of their actual usefulness.

Developing a Collaborative Online Environment for History – the Experience of British History Online

Blaney, Jonathan

jonathan.blaney@sas.ac.uk

Institute of Historical Research, UK

This paper will discuss the potential impact upon historical research of British History Online's annotation tool. British History Online (BHO) (www.british-history.ac.uk) is a digital library containing some of the core printed primary and secondary sources for the medieval and modern history of the British Isles. Created by the Institute of Historical Research (IHR), which is part of the University of London's School of Advanced Study, BHO is a rigorous academic resource used by researchers at postgraduate level and above. The IHR is centrally placed within academic history in the UK, and as such it is highly regarded within the profession.

Two years ago, as part of an Arts and Humanities Research Council (AHRC) grant, BHO undertook to provide an annotation tool alongside the digitisation of some 500 key historical sources for early modern history: the Calendars of State Papers. These calendars summarise the manuscript heritage of the working of the state in the early modern period. Essential research tools though these calendars are, they were mainly compiled in the Victorian period, and are known to be inadequate and erroneous in some cases. Furthermore, changes in perspective on history and subsequent research means that the calendars are badly in need of updating: papers which modern editors would think worthy of close attention are sometimes treated very cursorily. BHO's annotation tool encourages the community of scholars to update, enlarge upon, and correct the calendars, and even to supply fuller transcriptions of documents.

At the planning stage of the tool's development the team looked at other online tools for user commenting, ranging from scholarly collections such as the Digital Image Archive of Medieval Music to non-academic sites such as the blog-style Diary of Samuel Pepys. This paper will describe the reasons why we decided that an annotation tool was preferable to a wiki, the design process for the tool, and BHO's subsequent attempts to engage the academic community in online collaborative work: now that the AHRC project has finished (but the

annotation tool remains active) DH 2010 will be a good time to assess the successes and failures this aspect of the project and the lessons that might be applicable to other digital resources for historians. The IHR is now involved in a collaborative project, Connected Histories, and the lessons of BHO's annotation tool will have a direct bearing on how the front end for Connected Histories is designed.

The paper will also touch upon the issues of moderation of academic work, the role of citation within web 2.0, and the constraints we imposed on annotators in this regard, intellectual copyright and the RAE, and the broader question of how humanities research culture might change as web-based collaboration becomes the norm.

The IHR is currently addressing the question of how the research community within history might be mobilised to work together online, in more general ways, with European collaboration, semantic web research tools, and VREs. The paper will conclude by briefly placing these in the context of work already done on the annotation tool.

References

- Blaney, J. and Winters, J..** 'The British History Online digital library: a model for sustainability?'. *Bulletin of the Belgian Royal Historical Commission*.
- MacGregor, J., et al.** (2009). 'Revolutionary reading, evolutionary toolmaking: (Re)development of scholarly reading and annotation tools in response to an ever-changing scholarly climate, Implementing New Knowledge Environments'. University of Victoria, October 2009. <http://bit.ly/c5LV4R>.
- Connected Histories*. <http://www.connectedhistories.org> (accessed 25-03-2010).
- Diary of Samuel Pepys*. <http://www.pepysdiary.com> (accessed 25-03-2010).
- Digital Image Archive of Medieval Music*. <http://www.diamm.ac.uk> (accessed 25-03-2010).
- Diigo*. <http://www.diigo.com/> (accessed 26-03-2010).
- Horizon Project*. <http://www.nmc.org/horizon> (accessed 25-03-2010).
- Zotero*. <http://www.zotero.org/> (accessed 26-03-2010).

From Codework to Working Code: A Programmer's Approach to Digital Literacy

Bork, John

jrbork@wcn.org

University of Central Florida

What does it mean to be digitally literate? Obviously it entails a basic familiarity with commonly used technologies, so that one may navigate the technological life world that has permeated nearly every aspect of the human one. One aspect of this knowledge is the recognition of computer languages, communications protocols, syntactic forms, passages of program code, and command line arguments, even when they have been taken out of their operational context for use as literary and rhetorical devices. In addition to the infiltration of the abbreviated language of email and text messaging into mainstream print media, it is now also commonplace to encounter programming keywords, symbols, operators, indentation, and pagination entwined with natural, non-technical, mother tongue expressions. *Codework* is the term associated with the literary and rhetorical practice of mixing human and computer languages (Hayles, 2004; Raley, 2002; Cramer, 2008). Types of codework span from intentionally arranged constructions intended for human consumption that do not execute on any real computer system, to valid expressions in bona fide programming languages that are meaningful to both human and machine readers. Examples of the former include the work of Mez (Mary-Anne Breeze) and Talon Memmott, of the latter, the work of John Cayley and Grahan Harwood (Raley, 2002; Fuller, 2008). Rita Raley notes, however, that of the popular electronic literature of the early twenty first century, there is "less code per se than the language of code." In addition to its infusion for literary effect, program source code may be cited in scholarly texts like conventional citations to explain a point in an argument. Although it is more common to encounter screen shots of user interfaces, examples of working source code appear on occasion in humanities scholarship. This study will briefly consider why working code has been largely shunned in most academic discourse, and then identify the types and uses of bone fide code that do appear, or are beginning to appear, in humanities scholarship. Its goal is to suggest ways in which working code – understood both as code that *works*, and as the practice of *working code* – plays a crucial

role in facilitating digital literacy among social critics and humanities scholars, and demonstrate through a number of examples how this effect may be achieved.

The first argument in favor of studying computer code in the context of humanities scholarship can be drawn from N. Katherine Hayles' methodological tool of Media-Specific Analysis (MSA). Probing the differences between electronic and print media when considering the same term, such as hypertext, requires comprehension of the precise vocabulary of the electronic technologies involved. A second, more obvious argument comes from the growing disciplines of Software Studies and Critical Code Studies. If critical analysis of software systems is to reveal implicit social and cultural features, reading and writing program code must be a basic requirement of the discipline (Fuller, 2008; Mateas, 2005; Wardrip-Fruin, 2009). As the media theorist Friedrich Kittler points out, the very concept of what code is has undergone radical transformations from its early use by Roman emperors as cipher to a generic tag for the languages of machines and technological systems in general; "technology puts code into the practice of realities, that is to say: it encodes the world" (Kittler, 2008). Or, following the title of Lev Manovich's new, downloadable book, software takes command. Yet both Kittler and Manovich express ambivalence towards actually examining program code in scholarly work. A third argument, which will form the focus of this study, is reached by considering the phenomenon of *technological concretization* within computer systems and individual software applications. According to Andrew Feenberg, this term, articulated by Gilbert Simondon, describes the way "technology evolves through such elegant condensations aimed at achieving functional compatibilities" by designing products so that each part serves multiple purposes simultaneously (Feenberg, 1999). The problem is that, from the perspective of a mature technology, every design decision appears to have been made from neutral principles of efficiency and optimization, whereas historical studies reveal the interests and aspirations of multiple groups of actors intersecting in design decisions, so that the evolution of a product appears much more contingent and influenced by vested interests. The long history of such concretizations can be viewed like the variegated sedimentation in geological formations, so that, with careful study, the outline of a technological unconscious can be recovered. The hope is that, through discovering these concealed features of technological design, the unequal distribution of power among social groups can be remedied. Feenberg's project of democratic rationalization responds to the implicit oppression of excluded groups and values in technological systems by mobilizing workers, consumers, and volunteers to

make small inroads into the bureaucratic, industrial, corporate decision making.

For computer technology in particular, digital literacy is the critical skill for connecting humanities studies as an input to democratic rationalizations as an output. Working code replaces the psychoanalytic session for probing the technological unconscious to offer tactics for freeing the convention-bound knowledge worker and high tech consumer alike. Many theorists have already identified the free, open source software (FOSS) community as an active site for both in depth software studies and for rich examples of democratic rationalizations (Fuller, 2008; Yuill, 2008; Jesiek, 2003). Simon Yuill in particular elaborates the importance of revision control software for capturing and cataloging the history of changes in software projects. As a corollary to this point, it can be argued that concealed within these iterations of source code are the concretizations that make up the current, polished version of the program that is distributed for consumption by the end users, and from which the technological unconscious may be interpreted. However, even when they are freely available to peruse in public, web-accessible repositories, these histories are only visible to those who can understand the programming languages in which they are written. Therefore, it is imperative that humanities scholars who wish to critically examine computer technology for its social and cultural underpinnings include working code - as practicing programming - in their digital literacy curricula.

References

- Cramer, Florian** (2008). 'Language'. *Software Studies: A Lexicon*. Fuller, Matthew (ed.). Cambridge, Mass: The MIT Press, pp. 168-74.
- Feenberg, Andrew** (1999). *Questioning Technology*. New York: Routledge.
- Fuller, Matthew** (2008). 'Introduction'. *Software Studies: A Lexicon*. Fuller, Matthew (ed.). Cambridge, Mass: The MIT Press, pp. 1-13.
- Hayles, N. Katherine** (2004). 'Print is flat, code is deep: the importance of media-specific analysis'. *Poetics Today*. **25**(1): 67-90.
- Jesiek, Brent K.** (2003). 'Democratizing Software: Open Source, the Hacker Ethic, and Beyond'. *First Monday*. **8**(10) (accessed 5 October 2008).
- Kittler, Friedrich** (2008). 'Code'. *Software Studies: A Lexicon*. Fuller, Matthew (ed.). Cambridge, Mass: The MIT Press, pp. 40-7.

Mateas, Michael (2005). 'Procedural literacy: educating the new media practitioner'. *On The Horizon. Special Issue. Future of Games, Simulations and Interactive Media in Learning Contexts.* 13(1) (accessed 21 October 2009).

Raley, Rita (2002) (8 September 2002). 'Interferences: [Net.Writing] and the practice of codework'. *Electronic Book Review.* (accessed 7 October 2009).

Wardrip-Fruin, Noah (2009). *Expressive Processing: Digital Fictions, Computer Games, and Software Studies.* Cambridge, MA: The MIT Press.

Yuill, Simon (2008). 'Concurrent version system'. *Software Studies: A Lexicon.* Fuller, Matthew (ed.). Cambridge, Mass: The MIT Press, pp. 64-9.

Non-traditional Prosodic Features for Automated Phrase-Break Prediction

Brierley, Claire

cb5@bolton.ac.uk
University of Bolton, UK

Atwell, Eric

eric@comp.leeds.ac.uk
University of Leeds, UK

The goal of automatic phrase break prediction is to emulate human performance in terms of naturalness and intelligibility when assigning prosodic-syntactic boundaries to input text. Techniques can be deterministic or probabilistic; in either case, the problem is treated as a classification task and outputs from the model are evaluated against 'gold standard' phrase break annotations in the reference dataset or corpus. These annotations may represent intentions (of the speaker or writer) or perceptions (of the listener or reader) about alternating chunks and boundaries in the speech stream or in text, where the chunking bears some relationship to syntactic phrase structure but is thought to be simpler, shallower and flatter.

In this paper, we begin by reviewing methodologies and feature sets used in phrase break prediction. For example, a tried and tested *rule-based* method is to employ some form of 'chink-chunk' algorithm (Liberman and Church, 1992) which inserts a boundary after punctuation and whenever the input string matches the sequence: open-class or content word (chunk) immediately followed by closed-class or function word (chink), based on the principle that chinks initiate new prosodic phrases.

We discuss the limitations of using traditional features in the form of syntactic and text-based cues as boundary correlates, with illustrative experimental predictions from a shallow parser and evidence from the corpus. We then discuss the limitations of evaluating any phrase break model against a "gold standard" which itself only represents one phrasing variant for an utterance or text.

There is an emerging trend of leveraging real-world knowledge to improve performance in machine learning, including speech and language applications. Nevertheless, we have diagnosed a deficiency of *a priori* knowledge of *prosody* in the feature sets used for the phrase break prediction task. In contrast, a competent human reader is able to project holistic linguistic insights, including

projected prosody, onto text and to treat them as part of the input (Fodor, 2002). In this respect, multiple prosodic annotation tiers in the Aix-MARSEC corpus (Auran *et al.*, 2004) have been revelatory, since they capture the prosody implicit in text and currently absent in learning paradigms for phrase break models.

Insights such as: (i) the *transferability* of the chinks and chunks rule; plus (ii) the possibility of encoding a variety of prosodic phenomena (including rhythm and beats) in categorical labels (*cf.* the Aix-MARSEC corpus); plus (iii) an appreciation of prosodic variance gleaned from corpus evidence of alternative parsing and phrasing strategies, have informed the creation of ProPOSEL (Brierley and Atwell, 2008a; 2008b), a domain-independent prosodic annotation tool.

ProPOSEL is a **prosody** and **part-of-speech** English lexicon of 104,049 entry groups, which merges information from several widely-used lexical resources for corpus-based research in speech synthesis and speech recognition. Its record structure supplements word-form entries with syntactic annotations from four rival POS-tagging schemes, mapped to fields for: default open and closed-class word categories; syllable counts; two different phonetic transcription schemes; and lexical stress patterns, namely abstract representations of rhythmic structure (as in *201* for *disappear*, with secondary stress on the first syllable and primary stress on the final syllable).

We then contend that native English speakers may use certain sound patterns as *linguistic signs* for phrase breaks, having observed these same patterns at rhythmic junctures in poetry. We also contend that such signs can be extracted from canonical forms in the lexicon and presented as input features for the phrase break classifier in the same way that real-world knowledge of syntax is represented in POS tags; and that like content-function word status or punctuation, such features are domain-independent and can be projected onto *any* corpus. One such sound pattern is the subset of complex vowels, which we define as the eight diphthongs, plus the triphthongs, of Received Pronunciation (Roach, 2000: 21-24).

Finally, we test the correlation between pre-boundary lexical items bearing complex vowels and gold-standard phrase break annotations on different kinds of speech via the chi-squared statistic, to determine whether the perceived association is statistically significant or not. Our findings indicate that this correlation is extremely statistically significant: it is present in contemporary, formal, British English speech (Brierley and Atwell, 2009) and seventeenth century English verse (Brierley

and Atwell, 2010a); and it holds for spontaneous as well as read speech, and for multiple speakers (Brierley and Atwell, 2010b). We hypothesise that while complex vowels seem to constitute phrase break *signifiers* in English, this may translate to a subset of the vowel system in other languages.

References

- Auran, C., Bouzon, C. and Hirst, D.** (2004). 'The Aix-MARSEC Project: an Evolutive Database of Spoken British English'. *Proc. Speech Prosody*. 2004, pp. 561-564.
- Brierley, C. and Atwell, E.** (2008a). 'ProPOSEL: A Prosody and POS English Lexicon for Language Engineering'. *Proc. 6th Language Resources and Evaluation Conference*. LREC, 2008.
- Brierley, C. and Atwell, E.** (2008b). 'A Human-oriented Prosody and PoS English Lexicon for Machine Learning and NLP'. In *Proc. 22nd International Conference on Computational Linguistics*. Coling, 2008.
- Brierley, C. and Atwell, E.** (2009). 'Exploring Complex Vowels as Phrase Break Correlates in a Corpus of English Speech with ProPOSEL, a Prosody and PoS English Lexicon'. *Proc. INTERSPEECH'09*.
- Brierley, C. and Atwell, E.** (2010a). 'Holy Smoke: Vocalic Precursors of Phrase Breaks in Milton's Paradise Lost'. *Literary and Linguistic Computing*. **25(2)**.
- Brierley, C. and Atwell, E.** (2010b). 'Complex Vowels as Phrase Break Correlates in a Multi-Speaker Corpus of Spontaneous English Speech'. *Proc. Speech Prosody*, 2010 (Forthcoming).
- Fodor, J. D.** (2002). 'Psycholinguistics Cannot Escape Prosody'. *Proc. Speech Prosody*. 2002, pp. 83-90.
- Liberman, M. Y. and Church, K. W.** (1992). 'Text Analysis and Word Pronunciation in Text-to-Speech Synthesis'. *Advances in Speech Signal Processing*. Furui, S. and Sondhi, M. M. (ed.). New York: Marcel Dekker, Inc..
- Roach, P.** (2000). *Phonetics and Phonology: A Practical Course*. Cambridge: Cambridge University Press, 3rd Edition.

How Do You Visualize a Million Links?

Brown, Susan

susan.brown@ualberta.ca

University of Alberta and University of Guelph,
English

Antoniuk, Jeffery

jeffery.antoniuk@ualberta.ca

University of Alberta, Orlando Project

Bauer, Michael

bauer@uwo.ca

University of Western Ontario, Computer Science

Berberich, Jennifer

jenn_b19@hotmail.com

University of Western Ontario, Computer Science

Radzikowska, Milena

mradzikowska@gmail.com

Mount Royal College, Communications

Ruecker, Stan

sruecker@ualberta.ca

University of Alberta, Humanities Computing

Yung, Terence

terence.yung@zerom3.com

Mount Royal College, Communications

In the past quarter century, established methods of literary history have been severely contested. On the one hand, syncretic, single-author histories have become problematic as a result of a combination of the expanded literary canon and a range of theoretical challenges. On the other, a demand for historicized overviews that reflect the radical recent reshaping in all fields of literary study has produced large numbers of both collectively written histories and encyclopedias or companions. Literary history thus tends towards compilations in which specialists treat their particular fields, at the cost of integration or of coherence. Meanwhile, the primary materials are increasingly available in digital form, and literary historical scholarship itself is increasingly produced digitally, whether as versions of established forms such as journal articles, or in resources that invoke the potential for new kinds of analysis. Major digital initiatives over the past decades have focused almost exclusively on digital resource creation: the increasingly pressing question is how to use this expanding body of materials to its fullest potential.

In this project, we investigate how literary historical analysis can be extended using various forms

of visualization, using the experimental *Orlando* Project as our test bed. *Orlando: Women's Writing in the British Isles from the Beginnings to the Present* is recognized as the most extensive and detailed resource in its field and as a model for innovative scholarly resources. Composed of 1,200 critical biographies plus contextual and bibliographical materials, it is extensively encoded using an interpretive Extensible Markup Language (XML) tagset with more than 250 tags for everything from cultural influences, to relations with publishers, or use of genre or dialect. The content and the markup together provide a unique representation of a complex set of interrelations of people, texts, and contexts. These interrelations and their development through time are at the heart of literary inquiry, and having those relations embedded in the markup, and hence processable by computer, offers the opportunity to develop new forms of inquiry into, and representations of, literary history. Such new opportunities of scale are often invoked using Greg Crane's seminal question, "What can you do with a million books?" (2006).

We need to be able to ask big, complex questions while remaining grounded in particularities, and we need new ways of representing answers to those questions. This requires new tools for scholarly research that can access, investigate, and present new aspects of the human story and history. In this context, we contend that 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).

The *Orlando* encoding system, devised for digital rather than print textuality, facilitates collaboratively-authored research structured according to consistent principles. The encoding creates a degree of cross-referencing and textual inter-relation impossible with print scholarship—not simply hyperlinking but relating separate sections of scholarly text in ways unforeseen even by the authors of the sections. It represents a new approach to the integration of scholarly discourse, one which allows the integrating components to operate in

conjunction with, rather than in opposition to, historical specificity and detail (Brown et al. 2006c). However, the search-and-retrieval model of the current interface for *Orlando*, while user-friendly in that it resembles first-generation online research tools, cannot exploit this encoding to the fullest. Search interfaces only find what the user asks for, whereas visualization enables exploration and discovery of patterns and relationships that one might not be able to search for.

For instance, the current interface permits users to search for authors by the number of children they had, and thus to explore the relationship between literary production and reproduction. The quantity of material in *Orlando* makes it difficult to see overall patterns amongst the results. Recent experiments with the Mandala browser have demonstrated that visualization permits one to see both interesting anomalies (e.g. in lives which have demanded the use both of the childlessness tag and the children tag), and larger patterns, such as the non-correlation between high literary productivity and childlessness or small family size (Brown et al. 2008). These preliminary investigations confirm Moretti's argument (2005) that visual representations enable kinds of literary historical inquiry that are not supported by conventional search interfaces. *Orlando* has the added advantage of making it possible to dive back into the source material to see the specifics from which the representation is produced.

In addition to the Mandala experiments, we have also been working on a set of designs for visually summarizing relationships in a manner that allows interactive exploration (e.g. Fig. 1). Building on the large body of previous literature in network visualization (e.g. Barabási 2002; Watts 2003; Christakis and Fowler 2009), we are experimenting with new visual representations for networks of people.



Fig. 1: One of several concepts for summarizing relationships among authors in *Orlando*. Here the authors on the path

of connection are shown as coloured circles, where size is frequency and a unique colour is assigned to each author.

Interviews and observations of users at a recent hackfest provided some excellent insights into the sorts of interfaces that are likely to appeal to users wanting to explore embedded relationships in a body of texts in an open-ended way. While point-to-point visualization was considered to have some value, more excitement was generated by open-ended interface sketches, such as a visualization that resembles a cityscape, even when these were much less representational than conventional interfaces for the humanities.

At the same time, Brown and Bauer have been working on a visualization tool that illustrates the challenges facing the project of representing all of *Orlando*'s semantically interrelated data through a graphical representation based on nodes and edges. It highlights the difficulty of providing prospect when dealing with a large and complexly structured data set, since the full set of relationships even of a moderate subset of the 1200 writers becomes unreadable, with over 16 million edges in the graph. We have done some work to explore algorithms and interfaces to accommodate these large data spaces and multitudes of relationships and tags. The challenge is to provide the researcher with a means of perceiving or specifying subsets of data, extracting the relevant information, building the nodes and edges, and then providing means to navigate the vast number of nodes and edges, especially given the limited amount of space on a computer monitor. The figures below illustrate some of the aspects of the tool.



Fig. 2: several writers and their interconnections

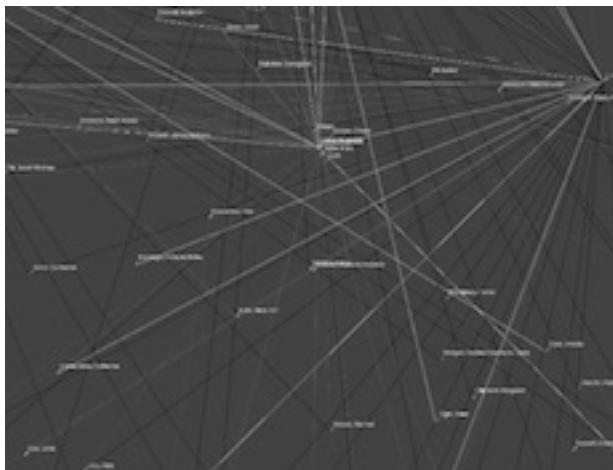


Fig. 3: zoomed view of relationships

The nodes (at the centre of the starbursts) represent writers, while the blue dots show other individuals (Fig. 2). The edges are shown as differently colored lines indicating different kinds of relationships as determined by tags (identified in the colored boxes). A researcher can display names, hide certain edges by deselecting tags, and zoom in and move around a large graph of nodes and edges (Fig. 3). The tool is a starting point for evaluating existing computational approaches and graphical displays of relationships as a means of exploring literary questions. It raises exciting questions regarding the integration of data mining approaches with a graphical interface, particularly for scholars suspicious of abstractions. Computationally, the question of how to make such a tool accessible to remote users is a challenge.

This paper will compare the various approaches to visualizing links that we have employed to date on this data, and reflect on them in relation to both the literature on visualization approaches and our user feedback as a means of advancing our thinking on the challenge of creating interfaces for exploring large numbers of interlinkages within or between humanities resources.

References

- Barabási, Albert-Lászlo** (2002). *Linked: The New Science of Networks*. Cambridge, MA: Perseus Publishing.
- Brown, Susan, Clements, Patricia, Grundy, Isobel, Ruecker, Stan, Antoniuk, Jeffery, Balazs, Sharon, Sinclair, Stéfan, Patey, Matt** (2008). 'Beyond Text: Using the Mandala Browser to Explore *Orlando*'. *Society for Digital Humanities (SDH/SEMI) Meeting. Congress of the Humanities and Social Sciences Federation of Canada*. University of British Columbia, June 2008.

Brown, Susan, Clements, Patricia, Grundy, Isobel (2006c). 'Scholarly Introduction'. *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*. Cambridge: Cambridge University Press. http://orlando.cambridge.org/public/svDocumentation?formname=t&d_id=ABOUTTHEPROJECT.

Crane, G. (2006). 'What do you do with a million books?'. *D-Lib magazine*. 3.

Drucker, J., Nowviskie, B. (2004). 'Speculative computing: Aesthetic provocations in humanities computing'. *A Companion to Digital Humanities*. Schreibman, Susan, Siemens, Ray, Unsworth, John (eds.). Oxford: Blackwell, pp. 431-447.

Christakis, Nicholas A., Fowler, James H. (2009). *Connected: The Surprising Power of Social Networks and How They Shape Our Lives*. New York: Little, Brown and Company.

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

Ramsay, Stephen (2003). 'Toward an Algorithmic Criticism'. *Literary and Linguistic Computing*. 2.

Unsworth, John (2005). 'New methods for humanities research'. *Lyman Award Lecture*. National Humanities Center Research Triangle Park, NC, 11 November 2005. <http://www3.isrl.illinois.edu/~unsworth/lyman.htm>.

Watts, Duncan J. (2003). *Six Degrees: The Science of a Connected Age*. New York: W. W. Norton & Company.

Digital Libraries of Scholarly Editions

Buchanan, George

g.r.buchanan@gmail.com

School of Informatics, City University, London

Bohata, Kirsti

K. Bohata@swansea.ac.uk

Centre for Research into the English Literature and Language of Wales (CREW), Swansea University

Digital libraries are a key technology for hosting large-scale collections of electronic literature. Since the first digital library (DL) systems in the early 1990s, the sophistication of DL software has continually developed. Today, systems such as DSpace and Greenstone are in use by institutions both large and small, providing thousands of collections of online material. However, there are limitations even to “state-of-the-art” DLs when considering digital humanities.

Contemporaneously with the growth of DL technology, digital scholarly editions of significant texts and archival material have emerged. In contrast to digital libraries, where there are a number of readily available generic software systems, critical editions are largely reliant on bespoke systems, which emphasise internal connections between multiple versions. Whilst useful editions are online, and are increasingly used in scholarly endeavour, digital scholarly editions suffer from “siloing”: each work becoming an island in the ocean of the web.

Like ‘physical’ libraries, digital libraries provide consistent support for discovering, reading and conserving documents in large collections. For scholarly editions, these features present a potential solution to “siloing”. Without trusted digital repositories, preservation and maintenance are endemic problems, and providing consistent experiences and unified workspaces across many sites (i.e. individual texts) is proving highly challenging. However, current DL systems lack critical features: they have too simple a model of documents, and lack scholarly apparatus.

Digital library systems can readily contain electronic forms of “traditional” critical editions in static forms where each work is a separate and indivisible document. However, search facilities cannot then reliably distinguish between commentary and the primary content. Using XML formats such as TEI can permit this distinction to be made, but only with extensive configuration. Furthermore, the reading

experience in such a DL is likely to fall far below scholars’ requirements of digital editions.

European initiatives, such as DARIAH, focus on facilitating access to existing scholarly editions, with longer-term aims of fostering standards and interoperability of data. This approach presumes that each existing site (and hence, typically, edition) remains autonomous, and remains a discrete entity, which is then aggregated through a centralised service. It also admits the absence of standardised, highly functional storage and publication systems. Furthermore, this approach has been attempted in “federated” DLs, with only limited success. In federated DLs, unless every member uses the same software configured in the same manner, the appearance of each library differs and – worse – preservation remains in the hands of individual sites, and cross-site services (e.g. search) can only operate at a very rudimentary level.

There have been projects to develop generic scholarly edition software, but success has been limited. Shillingsburg [Buchanan 2006] highlights a number of such systems up to the mid-2000s. Few of these initiatives engaged with computer science, and the software systems have proved hard to maintain.

Digital library systems provide a potential route for providing collections of digital scholarly editions. However, they are not yet an answer. When a digital edition supports discussion between scholars, grounded on and linked to the text, the standard DL infrastructure requires extensive modification. Data structures are required to capture and store scholarly discourse, relate each item of discourse in detail to part of a complex document structure, and provide this through a seamless and consistent user interface. Multiple structures and complex document relationships fit uneasily within current DL software [Buchanan *et al.* 2007, Rimmer *et al.* 2008]. For instance, most DL software requires or assumes that any collection of documents is homogenous in terms of the interior structure of each document. This simply cannot be true of a collection including – say – diaries, journals, letters and novels. We need software that provides DL collection support with the ability to provide for complex document structure.

1. Current Work

The goal of our research is to develop software that transcends the current limitations of DL systems in supporting digital scholarly editions for the humanities. Our intention is that in turn organisations and publishers who seek to provide series of critical material can build upon software that is scalable, systematically engineered and sustainable. This software will also support the necessary complexity of critical editions and possess

a rich apparatus to support contemporary digital practices, not simply digitised forms of practice from the print era. Whilst no single system is likely to provide all the requirements of all possible circumstances, our aim is to create software that can provide the technical core of any collection of critical editions, with a minimum of effort. Adapting the system for a specific need may require extensive work, but only for more unusual circumstances. This would bring us to a point comparable to the support that current DLs give for simpler texts and scholarly practices. For users of scholarly editions – i.e. the research community – the presence of a common infrastructure and the increased ease of working across sites will very likely increase research activity across multiple ‘editions’.

2. Context and Motivation

This project has identified Wales as presenting an interesting case study. It is a distinct cultural entity with an abundance of valuable written cultural material and a sizable scholarly community researching the cultural life and output of the nation. Reflecting the bilingual linguistic identity of the nation, there are extensive archives and printed matter in national, university, local government and private hands in both Welsh and English (as well as other languages). Wales suffers from a poor physical infrastructure, and this has motivated the provision of digital access to cultural material, from the early days of the National Library of Wales’s digitisation projects (e.g. ‘Campaign’ 1999) to the present.

While the National Library of Wales has done outstanding pioneering work in digitisation of its collections and remains an asset in our selection of Wales as a ‘case study’, their remit does not extend to the interpretation of their collections. Despite considerable demand from scholars in Wales and beyond for digital critical editions of Welsh material (in both languages) no one project has access to the technical expertise to create software that embodies the requirements of the scholarly community. The motivation of our project is to build a common infrastructure that both enables each project to produce high-quality scholarly work, and provides for consistent access and preservation of that work.

3. Understanding User Needs

To undertake this work requires not only technical expertise, but also a systematic study of the requirements of scholarly practice in the digital age. To date, we have reviewed the existing literature, and gained an initial set of requirements from a retrospective analysis of data from the recent User Centred Interactive Search (UCIS) project at University College London [Rimmer *et al.* 2008].

The UCIS project revealed that many technical difficulties emerged when configuring DL systems, even with relatively simple digital humanities material. Humanists do not necessarily search for material that directly corresponds to the “book” or “document” level of a particular library. Items may be sought that constitute part of a single document (e.g. a poem in a collection of poetry), and conversely larger works may be realised in several separate “documents”. Search and browse facilities typically work only at one level, typically consonant with either a book or article. However, collections are frequently heterogeneous and multi-layered. In the case of critical editions, the complexity of document structure and users’ tasks is even greater.

A second problem is that humanists often require different variants of one work. Though library infrastructures can relate these together, using standard features alone is insufficient [Shillingsburg 2006]. Even the more developed features that of a few DL systems are simplistic when compared to the complex relationship between different renditions and editions of a work that critical scholarship requires. Current methods relate entire separate items together – e.g. a chapter to a chapter – but scholarly criticism and annotation do not neatly conform to the clean structural boundaries favoured in computer or library science.

Thirdly, whilst some specific digital library installations do permit individual works to be linked to their author, or even specific words in a text to related material, this is not a standard part of DL software, and in contrast to the advanced facilities available in the best hypertext systems, the current technologies are primitive [Goose *et al.* 2000].

These shortcomings represent only a few of the problems already identified, and while we have developed partial solutions to parts of these, our technologies are not yet comprehensive, and other challenges have yet to be answered at all.

4. Summary

This presentation will articulate the shortcomings and problems raised when collections of critical materials are hosted through current digital library systems, and the contrasting “siloing” problems faced in the field of digital critical editions. We will demonstrate the requirements that will have to be matched to provide a single software system that delivers the needs of critical editions whilst also providing methods to develop collections of critical works. We illustrate how some of these requirements can be met, and prioritise and elucidate the remaining challenges in creating a unified system.

References

- Shillingsburg, Peter** (2006). *From Gutenberg to Google*. Cambridge: Cambridge University Press.
- Buchanan, George** (2006). 'FRBR: Enriching and Integrating Digital Libraries'. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Chapel Hill, NC, USA, June 11-15, 2006, pp. 260-269.
- Rimmer, Jon, Warwick, Claire, Blandford, Ann, Gow, Jeremy, Buchanan, George** (2008). 'An examination of the physical and the digital qualities of humanities research'. *Information Processing and Management*. Elsevier 3: 1374-1392.
- Buchanan, George, Gow, Jeremy, Blandford, Ann, Rimmer, Jon, Warwick, Claire** (2007). 'Representing aggregate works in the digital library'. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Vancouver, BC, Canada, June 17-23, 2007, pp. 247-256.
- Goose, Stuart, Hall, Wendy, Reich, Siegfried** (2000). 'Microcosm TNG: A Framework for Distributed Open Hypermedia'. *JIEEE MultiMedia*. 3: 52-60.

Digital Mediation of Modernist Literary Texts and their Documents

Byron, Mark

mark.byron@sydney.edu.au

University of Sydney

This paper will demonstrate an advanced work in progress, the digitised manuscript and transcription of Samuel Beckett's novel *Watt* (composed in 1941-45 and first published in 1953). Discussion of the project will centre upon the digital resources buttressing the presentation of manuscript material and a range of related analytic features, and will outline some of the more significant ways in which specifically digital treatment of the material opens up new lines of literary and textual analysis. Indeed, some foundational concepts of textuality come into sharp focus by virtue of digital treatment of textual materials. Some of these concerns will be illustrated by way of examples taken from the *Watt* project, and by a fuller view of the complex relationship between text and manuscript arising from the project.

1. SBDMP and the Digitised Edition of the *Watt* Notebooks

The digital and scholarly resources required to produce a digitised literary transcription are not trivial. Two related questions must frame any such project: what scholarly need is being met by the production of such an edition? What specific innovations are made available by virtue of its digital delivery?

The digital transcription of Beckett's *Watt* is an instalment of a larger international project – the Samuel Beckett Digital Manuscript Project – which aims to have all of Beckett's literary manuscripts transcribed and represented in digital form. This initiative responds to a profound deepening of scholarly interest in modernist manuscripts as potential sources of literary hermeneutic attention, and in concert with this focal shift, a renewed interest in theories of textuality and textual criticism. The specific (and heightened) relevance to this particular text in Beckett's oeuvre is immediately apparent on viewing the complex series of heavily revised and illustrated notebooks that constitute the manuscript of *Watt*. The primary materials do not lend themselves easily to conventional print publication, and indeed several dominant textual features would be lost or deeply submerged within

any codex structure. For example, the relationships between dispersed narrative episodes and fragments within the manuscripts cannot be represented adequately in the linear structure of the codex, nor the complex patterns of transmitted, dispersed and submerged material between the manuscript and the published editions of *Watt*.

Of all of Beckett's major texts, *Watt* has received the least critical attention, despite significant scholarly curiosity regarding the deep ambiguity of the published narrative and the baroque nature of its manuscript archive.¹ One reason for this oversight pertains directly to the digitised manuscript project: the materials extend to nearly a thousand pages of autograph manuscript in Beckett's notoriously difficult hand. Few scholars have read any of the primary materials, and only a very few have read them completely. The well-known hermeneutic difficulties presented by the published narrative are thus in no way adequately understood in relation to the primary materials, because they themselves constitute a kind of *terra nullius*. By representing and transcribing the manuscript archive of this pivotal text in digital form, such relations between the archive and publication can begin to proceed in an informed way, and more adequate editorial and hermeneutic strategies can be brought to bear on this most inscrutable of Beckett's texts.

The difficulties of reading Beckett's manuscript and text are, in part, aesthetic. The manuscript was composed during the Second World War, when its author was displaced in the south of France, at a time when reflections upon the efficacy of literary expression were most acute. In addition, the fragments, riddles, and non sequiturs in the published novel (first published in 1953) strongly imply a process of archivisation of fuller manuscript material, or more accurately, providing keys by which to unlock abundant manuscript contents. By providing coherent and searchable access to such a large and complex document, the digitised manuscript project provides the grounds for extensive investigation into hitherto inscrutable textual features in the published text, and provides space for reflection on variant narrative structures and the evolution of literary works more generally.

2. Digital Technology and Editorial Practice

The presence of digital technology in scholarship has become increasingly prominent in recent years. Digital aides to scholarship (online library catalogues, concordances, etc.) provide extensions to existing scholarly tools and practices, facilitating certain kinds of scholarship. Primary sources can be identified by means of web-based archive catalogues,

and online digital representations of manuscripts allow scholars to conduct particular kinds of work at geographical distance. Whilst access to the physical document may be desirable or even critical in the final event, several stages of a research project can be accomplished prior to such access. Digital extensions of traditional analogue research tools are perfectly commonplace in most disciplines, and (in theory) are not particularly difficult to integrate into a disciplinary mentality.

Recent innovative approaches to scholarly editing tend to imply or assert the relevance of a wider array of documentary sources: genetic editions seek to incorporate all manuscript material and published versions of a text, as well as a rationale of any stemmatic relationship between them, in an attempt to provide a "total" text; social text methods seek to integrate erstwhile secondary documents and materials into the very conceptual fabric of a text, as constituent parts of a text's identity. These more aggregative models of text identity, and more specifically the texts to which they pertain, are clearly conducive to presentation as digital scholarly editions. Conversely, digital modes of representing literary texts can bring questions of a text's identity into sharp focus. For example, the representation of multiple textual witnesses in collation software such as Juxta² or Versioning Machine³ alters rather profoundly the reader's apprehension of the textual matter at hand. The text is digitally mediated and may be represented by transposed digital reproductions and transcriptions suitably marked up for digital display. But this mediation can go to the very heart of what is considered to be the text.

Digital scholarly editions can do two things that seem fundamentally new: firstly, a potentially large corpus of material can be represented in one space, and manipulated in ways simply not possible in the world of physical manuscripts and codex editions: a basic premise of the digitised manuscript of *Watt*. Secondly, digital collations allow for manipulations of the text material that are visually straightforward and intuitively intelligible, whilst bearing profound implications for the text's identity and the authority of textual evidence. The digital manuscript of *Watt* deploys an interface powered by the Apache Tomcat servlet container, which represents files marked up in XML, in a streamlined version of the TEI5 protocols. A high-resolution digital image of the manuscript page appears alongside the marked-up transcription and attendant tools for analysing the transcribed document. In the case of this particular project, the use of Juxta collation software is not a straightforward choice, given that the manuscripts accord very closely to the published text in many places but diverge almost absolutely in many others. The relationship between text and archive is by no

means self-evident, or even chronologically linear, witnessed by the density of cryptic allusions and riddles in the published narrative: many of these may only be understood following a close reading of the more expansive manuscripts episodes from which they are sedimented.

The application of the Juxta software to such an editorial project as the digital variorum edition of Ezra Pound's *Cantos* offers an instructive counterpoint, providing a view of the way in which a well-developed and intuitively graspable digital aid offers new opportunities for new documentary and analytic research. These two examples provide one aspect by which to view the question of digital tools for literary research: does each project in the field of digital humanities require custom digital tools, or are there ways to engineer convergences that continue to provide each project with the specific resources it needs? This remains an open question, inspiring in equal parts an anxiety of resource-intense customisation, on the one hand, and the very exciting prospect of powerful convergences of digital tools in literary research on the other. One critical implication for literary studies is that wherever this question may lead, the nature and status of text identity will demand radical investigation.

3. The Digital-Textual-Literary Future

Recent advances in digital scholarly tools present exciting possibilities for scholarship and for reconsiderations of the paradigms of scholarship. They also present a basic challenge to the work undertaken in literary studies, by calling into question some of the most fundamental conceptual paradigms. The opportunity exists for significant developments in the theory of textual criticism. Whilst it is critical not to overstate the kinds of change made possible by digital scholarship – some apparent paradigm shifts are simply incremental changes to concepts and methods that remain integral to literary scholarship – it seems clear that we are only beginning to understand just what may be possible in the digital domain.

References

- Coetzee, J.M** (1972). 'The Manuscript Revisions of Beckett's "Watt"'. *JML*. **2,4**: 472-480.
- Hayman, D** (1997). 'Beckett's "Watt" – the Graphic Accompaniment: Marginalia in the Manuscripts'. *Word & Image*. **13.2**: 172-182.
- Hayman, D.** (1999). 'Nor Do My Doodles More Sagaciously: Beckett Illustrating "Watt"'. *Samuel*

Beckett and the Arts: Music, Visual Arts, and Non-Print Media. Oppenheim, L. (ed.). New York and London: Garland, pp. 199-215.

Kennedy, S. (1998). "'Astride of the Grave and a Difficult Birth': Samuel Beckett's "Watt" Struggles to Life". *Dalhousie French Studies*. **42**: 115-147.

Pilling, J. (1994). 'Beckett's English Fiction'. *The Cambridge Companion to Beckett*. Pilling, J. (ed.). Cambridge: Cambridge University Press, pp. 17-42.

Notes

1. J. M. Coetzee described the *Watt* manuscript material and hypothesised its stages of composition in his PhD dissertation over thirty years ago at the University of Texas at Austin. An epitome of this description and analysis was published in his essay, "The Manuscript Revisions of Beckett's *Watt*," *JML* 2.4 (1972): 472-480. Other discussions include: Sighle Kennedy, "'Astride of the Grave and a Difficult Birth': Samuel Beckett's *Watt* Struggles to Life," *Dalhousie French Studies* 42 (1998): 115-147; David Hayman, "Beckett's *Watt* – the Graphic Accompaniment: Marginalia in the Manuscripts," *Word & Image* 13.2 (1997): 172-182 and "Nor Do My Doodles More Sagaciously: Beckett Illustrating *Watt*," in Lois Oppenheim, ed., *Samuel Beckett and the Arts: Music, Visual Arts, and Non-Print Media* (New York and London: Garland, 1999), pp. 199-215; and John Pilling, "Beckett's English Fiction," in Pilling, ed., *The Cambridge Companion to Beckett* (Cambridge: Cambridge UP, 1994), pp. 17-42.
2. Juxta was originally developed as a collation tool for Jerome J. McGann's digital Rossetti Archive <<http://www.rossettiarchive.org/>> and is now housed under the auspices of the Institute for Advanced Technology in the Humanities and NINES (a digital research environment for nineteenth century studies), Alderman Library, University of Virginia.
3. Susan Schreibman began developing Versioning Machine <<http://v-machine.org/>> in 2000. It is housed at the University of Maryland Libraries and the Maryland Institute for Technology in the Humanities.

Detection of Citations and Textual Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project

Büchler, Marco

mbuechler@eaqua.net

Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany

Geßner, Annette

agessner@eaqua.net

Ancient Greek Philology Group, Institute of Classical Philology and Comparative Studies, University of Leipzig, Germany

Heyer, Gerhard

gheyer@eaqua.net

Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany

Eckart, Thomas

teckart@eaqua.net

Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany

"Users of this or any edition are warned that the textual variants presented by citations from Plato in later literature have not yet been as fully investigated as is desirable". This shortcoming, characterized by Kenneth Dover (Dover, 1980) is still existent and is unlikely to be corrected quickly by traditional research techniques. Textual reuse plays an important role in Classical Studies research. Similar to modern publications, classical authors used the texts of others as sources for their own work. In ancient texts, however, a less stronger form of word by word citation can be observed. Additionally, the complexity of ancient resources disallows fully manual research.

From a bird's eye view there are different points of view to the problem of textual reuse implying different research interests (Büchler and Geßner, 2009):

- A **Computer Science** perspective focuses on algorithms (*technical view*): Which algorithm is better than others? The scope of this research is wide ranging and also relates to plagiarism

detection in modern texts like theses at universities (Potthast et al., 2009).

- A **Historian** is interested in more complex correlations (*macro view*). For this kind of work a dedicated user interface is necessary to figure out relations between e.g. chapters of a book and their citation usage on a timeline.
- The research interests of a **Classical Philologist** focus on the textual differences between the original text and its variants in citations (*micro view*). These varying requirements necessitate designing different user interfaces for these three kinds of researchers.

Within the eAQUA project we are investigating the reception of Plato as a case study of textual reuse in ancient Greek texts. Our research is carried out in two steps. On the *technical level*, we firstly extract word by word citations. This is achieved by combining syntactical ngram overlappings (Hose, 2009 and Büchler, 2008) and significant terms for several of Plato's works. In the second step the constraints on syntactic word order are relaxed. This is done by combining text mining and information retrieval techniques. A graph based approach is then introduced that can deal with free word order citations. The key concept is not syntactically based, but focuses on the semantic level to extract the relevant *core information* of a used citation. Then the information is represented as a formal graph that is similar to the *Lexical Chaining* approach (Waltinger et al. 2008) that is often used for text summarisation (Yu et al. 2007). On the one hand syntactical and semantic approaches are only used to select reuse candidates with a small set of uncommon matching words within a citation. On the other hand, a complete pairwise comparison of all of the nearly 5.5 million sentences in the TLG corpus would require approximately 1000 years due to the squared complexity of $O(n^2)$ that was used for example to compare the Dead Sea Scrolls with the Hebrew Bible (Hose, 2009). For this reason, an intelligent pre-clustering of relevant reuse candidates is needed. Such a divide and conquer strategy reduces the complexity dramatically. Whilst the second step only increases the degree of free word order, in the third step the algorithm is expanded by similarly used words like *go* and *walk*. Those candidates are computed by similar cooccurrence profiles. The three levels briefly described above are only one dimension of reuse exploration. Other relevant dimensions that will be discussed are the *degree of preprocessing* as well as the *visualisation* of textual reuse in terms of citations.

In the field of preprocessing the main focus lies on *tokenisation* (more active tokenisation is needed with ancient texts than on modern languages),

normalisation (reducing all words internally to a lower-case representation without diacritics) and *lemmatisation* (reducing all words internally to a word's base form). This dimension can speed up the algorithm and also improves the results for strongly inflected languages like Ancient Greek.

Leaving the technical point of view of computer scientists, the research of Classicists includes both an application of a *macro view* for Historians as well as one for the *micro view* of Classical Philologists. The visualisation dimension of textual reuse is important since text mining approaches typically generate a huge amount of data that can't be explored manually. This is shown in Fig. 1. Whilst the light grey area marks Neoplatonism (about 5. AC) the grey ranges highlight Middle Platonism (about 2. AC). Taking Plato's *Timaeus*, one can clearly identify that both phases of Plato's reception (see Fig. 1 – top) are based on different "chapters" of *Timaeus* (bottom).

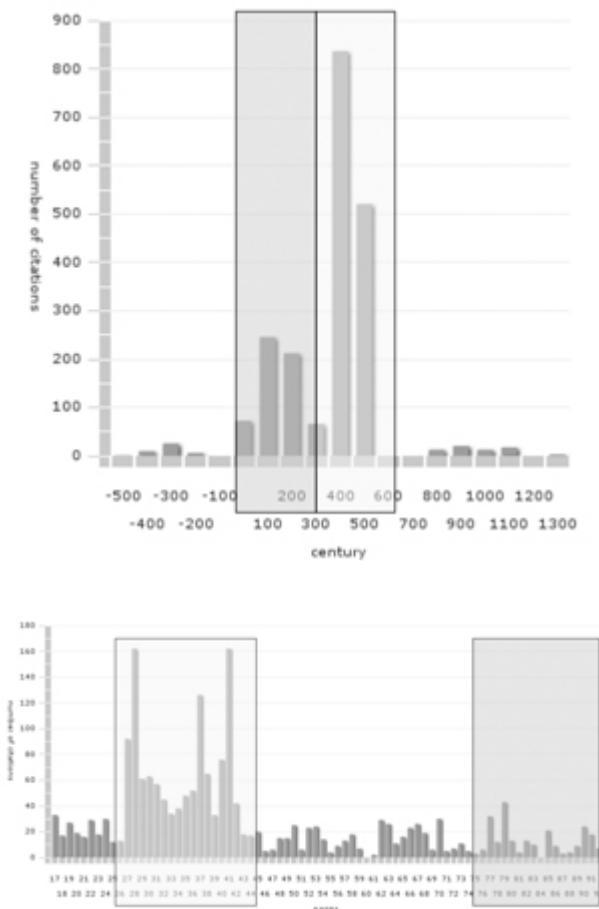


Fig. 1. Macro view: Screen of an interactive visualisation for citation usage. Citation distribution by Stephanus pages of Plato's *Timaeus*. The highest peak of the first picture is strongly correlated with the citation usage of the pages 27 to 42 of the second picture: Neo Platonism.

As Fig. 1 is of stronger interest for Historians, there is also a requirement for a visualisation for researchers from the field of Classical Greek Philology. As shown in Fig. 2, a visualisation highlighting the differences

in citation usage is necessary. This is especially important if longer citations are investigated.

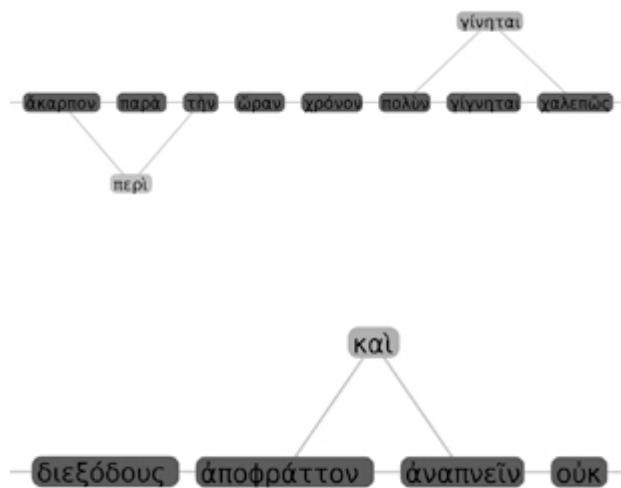


Fig. 2. Micro view: Highlighted differences of citations (green, orange) in relation to original text of Plato (blue).
Top: The orange word highlights the same word but including a language evolution of about 10 centuries.
Bottom: An included word (orange) in the citation is shown.

Additionally, it will be demonstrated how to detect different editions of the same original text. Such completely unsupervised approaches are important to investigate the scientific landscape of text digitisation. Furthermore, the relation to modern plagiarism detection will be given as well as the importance of building modern representative corpora since especially web corpora typically contain several duplicates of the same text.

In the evaluation section different results related to the comparison of various approaches on several text genres will be shown. An example of those results is given by contrasting citations of Plato's work with the textual reuse of the Attidographers. Whilst citations of Plato can be extracted quite well by the syntactical approach even with very low similarity thresholds, the same approach works with an accuracy smaller than 20% for textual reuse of the Attidographers.

Finally, results of a still in progress manual evaluation will be presented relating to the question of how and why a passage was cited.

References

- Büchler, M. (2008). *Medusa: Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung*. Saarbrücken: VDM Verlag Dr. Müller.

Büchler, M. and Geßner, A. (2009). 'Citation Detection and Textual Reuse on Ancient Greek texts'. *2009 Chicago Colloquium on Digital Humanities and Computer Science*. Argamon, S. (ed.). Chicago.

Dover, K. (1980). *Plato: Symposium*. Cambridge: Cambridge University Press.

Hose, R. (2009). *CS490 Final Report: Investigation of Sentence Level Text Reuse Algorithms*. <http://www.cs.cornell.edu/BOOM/2004sp/ProjectArch/DeadSea/index.html> (accessed 29 October 2009).

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A. and Rosso, P. (2009). 'Overview of the 1st International Competition on Plagiarism Detection'. *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN09)*. Stein, B., Rosso, P., Stamatatos, E. Koppel, M. and Agirre, E. (ed.). CEUR-WS.org, pp. 1-9.

Waltinger, U., Mehler, A. und Heyer, G. (2008). 'Towards Automatic Content Tagging: Enhanced Web Services in Digital Libraries Using Lexical Chaining'. *4th Int. Conf. on Web Information Systems and Technologies (WEBIST '08), 4-7 May, Funchal, Portugal*. Cordeiro, J. and Filipe, J. and Hammoudi, S. (ed.). Barcelona: INSTICC Press, pp. 231-236.

Yu L., Ma, J., Ren, F. and Kuroiwa, S. (2007). 'Automatic Text Summarization Based on Lexical Chains and Structural Features'. *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)*. V. 2, pp. 574-578.

No Representation Without Taxonomies: Specifying Senses of Key Terms in Digital Humanities

Caton, Paul

pncaton@gmail.com
INKE Project

INKE Research Group

inke.project@gmail.com
INKE Project

Digital humanities practitioners typically deal with polysemous terms by specifying the intended sense of a term in accompanying documentation (when it is one of the set of terms in a schema) or by giving a localized qualification (when the term is being used in a scholarly article). Granted, practitioners do interrogate their use of ubiquitous terms: 'theory,' 'model,' and 'text,' for example, have all been critically examined.¹ These discussions, however, have not visibly affected the prevailing ad hoc, localized approach to sense disambiguation.

In ordinary language use multiple senses are the norm: we might hope for greater precision in an academic field, but cannot assume it. "After all," writes Allen Renear apropos of conflicting views on the essential characteristics of textuality, "there is not even a univocal sense of 'text' within literary studies: Barthes's 'text' can hardly be Tanselle's 'text'" ("Out" Note 1 124). The more finely senses are distinguished, though, the greater the need for documentation to point to, the greater the amount of documentation there must be, and the greater the requirement that digital resources make all the necessary pointers available.

There is a case, then, for relieving the polysemous burden carried by terms like 'text'. This could be done either by shifting some senses onto different terms or by adding an agreed upon set of clearly defined qualifiers to the original term. One example of different terms being available is the FRBR Group One entity types (IFLA Study Group 3.2). It may not have been the *intention* of the IFLA Study Group to provide alternatives for 'text', but unquestionably each Group 1 entity type - work, expression, manifestation, and item - corresponds to an existing sense of 'text' and can therefore be used in place of it. However, while these types do capture some broad distinctions, the set is very small.

More ambitious is the taxonomy of texts proposed by Shillingsburg as part of his overall concept of a 'script act.'² Here the semantic burden is shifted to a qualifying phrase and 'text' has the constant sense of a sign sequence (in material or immaterial state), whose existence is established by at least one material instantiation, and which is intended as a unitary communication (whether actually finished or not). Extrapolating from this, we can say that--in relation to this taxonomy--'textuality' is the exhibiting of such properties, and 'text' as a general phenomenon (that is, as a mass noun rather than a count noun) is some quantity of that which exhibits 'textuality'.

These definitions are ours and not Shillingsburg's, but derive from his definitions and are consistent with the principles upon which his taxonomy is based. Furthermore, they accord with common senses of those terms. We emphasize this both because it has methodological implications and because it helps us rethink a notion of 'text' that is well-known in the digital humanities community and to see its proper relation to the senses just described.

The quote from Renear given earlier comes from his discussion of "theory about the nature of text" coming out of the electronic text processing and text encoding localities ("Out" 107). The view Renear himself espouses--"Pluralism"--developed as a refinement of the earlier view--"Platonism"--associated with the assertions made by de Rose et al in the paper "What is Text, Really?" This line of thinking has presented itself as *definitional*, offering a sense to associate with 'text.' Also, by emphasizing its origins in work on automated document processing, it presents this sense of 'text' as *fundamental*: that is, a more universal sense of 'text' than any sense coming from the traditional humanities localities, because it is as applicable to tax forms, memos, and technical manuals as to novels, plays, and poems. The third thing to note is that this approach has used 'text' in both mass noun and count noun senses interchangeably, and so whatever is said about one applies equally to the other. In the Pluralist view, what defines text is the presence of one or more structures of content objects. We believe this view actually has the opposite effect of what it originally intended because, despite its avowedly universal scope, it actually imposes a greater restriction on what qualifies as a text than Shillingsburg's taxonomy does. Shillingsburg's categories have the form QualifyingLabel+'text', where 'text' has the sense of a sign sequence as described earlier. The sentence "Call me Ishmael." clearly counts as 'text' in Shillingsburg's sense, and equally clearly does not count as 'text' in the Pluralist sense - unless we dilute the sense of the phrase 'content object' until it includes standard linguistic structural units such as

the clause, in which case the Pluralist sense simply becomes the same as Shillingsburg's sense.

What that line of thinking about text, texts, and textuality that runs from "What is Text, Really?" through "Out of Praxis" actually describes is a property that many--indeed most--texts exhibit, but that is not an *essential* property of a text. In a footnote to the discussion in "Out of Praxis" Renear acknowledges that the various meanings 'text' has in the various disciplinary localities do share a common ground, namely that "they all are efforts to understand textual communication." But he continues "I think that taxonomies of sense are best deferred until after we have a better understanding of actual theory and practice" (124). We think the conceptual help afforded by the clarity of Shillingsburg's distinctions shows the opposite is true: having taxonomies in place first betters our theoretical understanding.

That last statement brings out the 'chicken and egg' nature of this problem with terminology, as many scholars would doubtless argue that specifying a taxonomy like Shillingsburg's *presupposes* one's holding to a particular theory of text/textuality. Debating that, however, would in turn be helped by having a taxonomy of 'theory' available, because what that term means in digital humanities is itself hotly contested.

As helpful as we believe Shillingsburg's taxonomy to be, it only clarifies a few items of the "essential vocabulary," and while we think his overall 'script act' framework a good place to start, it needs adding to--for example, in the area that Shillingsburg calls "reception performance" (*Resisting* 77-80). Though he emphasizes his debt to McGann he doesn't attempt a taxonomy of the bibliographic codes that McGann considers such an important feature of production texts (*Textual* *passim*). Nor does he really say what happens to the notion of illocutionary point when we move from speech act to script act.³ This is work still to be done.

References

- Caton, Paul** (2003). 'Theory in Text Encoding'. *ACH/ALLC Annual Conference*. University of Georgia, Athens, Georgia, May 2003.
- Caton, Paul** (2004). *Text Encoding, Theory, and English: A Critical Relation*. Dissertation. Providence, RI: Brown University.
- DeRose, Steven J., et al.** (1990). 'What is Text, Really?'. *Journal of Computing in Higher Education*. 2: 3-26.

Eggert, Paul (2005). 'Text-Encoding, Theories of the Text, and the Work-Site'. *Literary and Linguistic Computing*. **4**: 425-435.

IFLA Study Group on the Functional Requirements for Bibliographic Records (2009). *Functional Requirements for Bibliographic Records: Final Report*. International Federation of Library Associations and Institutions, Amended and corrected version.

McCarty, Willard (2009). *Humanities Computing*. Basingstoke, Hampshire: Palgrave Macmillan.

McGann, Jerome (1988). *Social Values and Poetic Acts: The Historical Judgment of Literary Work*. Cambridge, Mass.: Harvard University Press.

McGann, Jerome (1991). *The Textual Condition*. Princeton Studies in Culture/Power/History. Princeton, New Jersey: Princeton University Press

Renear, Allen (1997). 'Out of Praxis: Three (Meta)Theories of Textuality'. *Electronic Text: Investigations in Method and Theory*. Sutherland, Kathryn (ed.). Oxford: Clarendon Press, pp. 107-126.

Renear, Allen (1997). 'Theory Restored: A Response to Caton'. *ACH/ALLC Annual Conference*. University of Gothenburg, Gothenburg, Sweden, June 2004.

Renear, Allen, Durand, David, Mylonas, Elli (1996). 'Refining our Notion of What Text Really Is'. *Research in Humanities Computing*. Ide, Nancy, Hockey, Susan (eds.). Oxford: Oxford University Press.

Robinson, Peter (2009). 'What Text Really Is Not, and Why Editors Have to Learn to Swim'. *Literary and Linguistic Computing*. **1**: 41-52.

Searle, John (1979). *Expression and Meaning*. Cambridge: Cambridge University Press.

Shillingsburg, Peter (2006). *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge: Cambridge University Press.

Shillingsburg, Peter (1997). *Resisting Texts: Authority and Submission in Constructions of Meaning*. Ann Arbor, Michigan: University of Michigan Press.

Shillingsburg, Peter (1991). 'Text as Matter, Concept, and Action'. *Studies in Bibliography*. **44**: 32-83.

Tanselle, G. Thomas (1995). 'The varieties of scholarly editing'. *Scholarly Editing: A Guide to Research*. Greetham, D.C. (ed.). New York: The Modern Language Association of America.

Notes

1. As a representative selection from the existing literature, see Caton "Theory"; Caton "Text Encoding", *passim*; DeRose et al; Eggert; McCarty, *passim*; Robinson; Renear "Out of Praxis"; Renear "Theory Restored"; Renear, Durand, and Mylonas.
2. See *Resisting* ch. 3. This is a revised version of his "Text" where the term originally used was 'write act.'
3. On illocutionary point see Searle 2. We find no treatment of it by Shillingsburg in either *Resisting* or *From Gutenberg*.

Modes of Seeing: Case Studies on the Use of Digitized Photographic Archives

Conway, Paul

pconway@umich.edu
University of Michigan

Digital humanities scholarship has expanded beyond its deep foundations in text analysis to find new meaning and knowledge through the creative reuse of historical photographs and other visual resources. Visual studies scholars in the humanities who wish to work primarily in the digital domain face a fundamental dilemma in the choice either to create “purpose-built” thematic collections tailored to specific studies (Palmer, 2008) or to make use of collections digitized for general purposes by an archives, a library, or other cultural heritage organization. “General-purpose” digital library collections are simultaneously mechanisms for delivering digital surrogates of archival holdings and new archival collections in their own right that reflect the decisions that digital curators make throughout the digitization process (Ross, 2007; Conway, 2008). The research issues associated with the actual use in humanities contexts of these large-scale general-purpose collections of digital images are profound and as yet largely unexplored (Saracevic, 2004). Concluding an important study establishing a typology of use in image retrieval, Graham (2004, p. 324) observes that “these uses do not tell us what was actually done with the images once they had been found.”

This paper reports on a multi-case study of the use of general purpose digitized photographic archives. The paper’s title is a play on John Berger’s somewhat forgotten pre-digital argument in *Ways of Seeing* that reproductions transform art (including of photographs and other graphical materials) into information, and in doing so expose original material objects to new uses not imagined by either the artist or, especially, the museums and archives that collect these artifacts. “It is not a question of reproduction failing to reproduce certain aspects of an image faithfully; it is a question of reproduction making it possible, even inevitable, that an image will be used for many different purposes and that the reproduced image, unlike the original work, can lend itself to them all” (Berger, 1972, p. 24). In suggesting that the post-modern critique has outlived its usefulness in the arena of visual studies, Mitchell calls for

moving beyond Walter Benjamin’s skepticism of the reproduction by embracing digital image surrogacy as superior. “In a world where the very idea of the unique original seems a merely nominal or legal fiction, the copy has every chance of being an improvement or enhancement of whatever counts as the original” (Mitchell, 2003, p. 487).

Efforts to extract evidence and meaning from the digitized photographic image extend well beyond the disciplines of art and art history to encompass history, a range of other social sciences, and increasingly the humanities. Humanists with a propensity toward visual studies run the gamut from skepticism to enthusiasm about the processes that digitally transform the material properties of original photographs and camera negatives. Koltun (1999, p. 124) claims that a digitized photograph “leaves behind another originating document whose disposal or retention can inspire other archival debates focused around original attributes and meanings not ‘translated’ into, even distorted by, the new medium.” Sassoon (2004, p. 199) largely sees diminished meaning (“an ephemeral ghost”) through digitization, whereas Cameron (2007, p. 67) projects archival properties onto the “historical digital object” that are distinctive and original. Skeptics and enthusiasts on both sides of this argument stake their claims with little regard for the actual uses of digitized historical photographs.

This paper exposes varying perspectives on “modes of seeing” by synthesizing case studies of seven deeply experienced researchers both within and outside the academy, ranging from scholars to serious avocational users to people whose livelihood depends on finding and using high quality representations of historical photographs. The group of study participants is broadly (but not statistically) representative of the variety of sophisticated humanities-oriented uses to which general purpose collections are put. The participants in the seven case studies vary widely in terms of demographic characteristics. Three are female; four are male. Their ages range from 30 to 67. The participants work and live east of the Mississippi River in five separate communities. Each case study revolves around a specific tangible product that was in some stage of completeness at the time of the interviews. The form of the products ranged from books and a dissertation, a complex and dynamic website, to a database for a membership organization. For their projects, participants made use of digitized photographs delivered from either the Library of Congress’s American Memory collection or the online catalog of the Prints and Photographs Division. Each of the five collections consulted is discrete within its particular delivery system. The Civil War Photographs collection is available through

interfaces to both the American Memory and the PPD databases. A 1872 Turkestan photographic album and photographs from the National Child Labor Committee are available in digital form only through the online catalog. Portions of the extensive Farm Security Administration/Office of War Information collection are distributed through the American Memory interface, but the entire digitized collection is fully available only through the online catalog. Finally, the Bain photograph collection, including a sizable sub-collection on American baseball, is fully available digitally through the online catalog and selectively through the American Memory interface.

The paper frames the findings on the use of digitized photographs in digital humanities scholarship within new theoretical perspectives on visual literacy (Elkins, 2008) and remediation (Bolter and Grusin, 1996), and the practical aspects of imaging for humanities scholarship (Deegan and Tanner, 2002). The case studies are constructed using an innovative multi-method qualitative approach that encompasses archival research in photo archives combined with a two-stage qualitative investigation. Stage one gathers background information and an assessment of expertise from interview subjects. Stage two consists of in-depth, semi-structured, in-situ interviews and observations. A three-part “thinking out loud” protocol extracts extensive commentary on the nature of individual and community expertise, on macro decision making strategies for creating the research product, and the character of micro-decisions on the choice and use of individual photographs. The descriptive evidence of “modes of seeing” is derived from a “grounded theory” analysis of interview transcripts.

Using extracted quotations and extensive visual examples, the paper presents an original typological model on the ways that perspectives of users on visual content, archival properties, and technical characteristics of digitized photographic archives combine to produce distinctive, but often intersecting “modes of seeing.” One mode, “Discovering,” takes maximum advantage of the visual detail discernable in high-resolution digital copies of camera negatives to find and contextualize new knowledge. A second mode, “Storytelling,” has a point of departure in the emotion evoked by wholly composed photographic images, seeking hidden narratives surrounding the subjects of the images, much in the way that textual archives yield their stories through the power of provenance. A third mode, “Landscaping,” finds meaning through the geospatial and temporal contexts of the images and the circumstances of their existence, sometimes providing a portal on technologically mediated power relations. All three modes carry either a “materialist” or “anti-materialist” stance that circumscribes the intimacy of

original source and digital surrogate. The two stances have much to say about trust, integrity, and the archival nature of digital collections for humanities scholarship.

The findings have at least three important implications for the use of general purpose collections of digitized photographs in a digital humanities context. First, the study demonstrates the relationship between the technical characteristics of digitally transformed photographs and the construction of visual narrative. Second, the study exposes how hidden archival properties embedded in the transformed archival photographic record create the context of use for scholars who privilege digital surrogacy over the material nature of original sources. Third, the study’s model of “modes of seeing” diversifies our understanding of how humanists interpret the visual order on, beneath, and beyond the visual plane of the photographic object.

Funding

This work was supported by the U. S. National Science Foundation [IIS-0733279]. Ricardo Punzalan provided valuable assistance in conducting the phase-one interviews.

References

- Berger, J., et al.** (1972). *Ways of Seeing*. London: Penguin Books and British Broadcasting Corporation.
- Bolter, J., Grusin, R.** (1996). 'Remediation'. *Configurations*. 4 (3): 311-358.
- Cameron, F.** (2007). 'Beyond the Cult of the Replicant: Museums and Historical Digital Objects – Traditional Concerns, New Discourses'. *Theorizing Digital Cultural Heritage: A Critical Discourse*. Cameron, F., Kenderdine, S. (eds.). Cambridge, MA: MIT Press, pp. 49-75.
- Conway, P.** (2009). 'Building Meaning in Digitized Photographs'. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science* 1. Chicago. <http://https://letterpress.uchicago.edu/index.php/jdhcs/article/viewFile/12/61> (accessed 3 March 2010).
- Deegan, M., Tanner, S.** (2002). *Digital Futures: Strategies for the Information Age*. London: Library Association Publishing.
- Elkins, J. (ed.)** (2008). *Visual Literacy*. New York: Routledge.
- Graham, M.** (2004). 'Enhancing Visual Resources for Searching and Retrieval'. *Literary and Linguistic Computing*. 3: 321-333.

Koltun, L. (1999). 'The Promise and Threat of Digital Options in an Archival Age'. *Archivaria*. **47** (Spring): 114-135.

Mitchell, W.J.T. (2003). 'The Work of Art in the Age of Biocybernetic Reproduction'. *Modernism/modernity*. **10** (3): 481-500.

Palmer, C. (2004). 'Thematic Research Collections'. *A Companion to Digital Humanities*. Schreibman, S., Siemens, R., Unsworth, J. (eds.). London: Blackwell, pp. 348-365.

Ross, S. (2007). 'Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries'. *Keynote Address at the 11th European Conference on Digital Libraries (ECDL)*. Budapest, 17 September 2007. http://www.ecdl2007.org/Keynote_ECDL2007_SROSS.pdf (accessed 3 March 2010).

Saracevic, T. (2004). 'How Were Digital Libraries Evaluated?'. *DELOS WP7 Workshop on the Evaluation of Digital Libraries*. Padova, Italy, 4-5 October 2004. http://comminfo.rutgers.edu/~tefko/DL_evaluation_LIDA.pdf (accessed 3 March 2010).

Sassoon, J. (2004). 'Photographic Materiality in the Age of Digital Reproduction'. *Photographs, Objects, Histories*. Edwards, E. (ed.). London: Routledge, pp. 196-212.

Digital Humanities Internships: Creating a Model iSchool-Digital Humanities Center Partnership

Conway, Paul

pconway@umich.edu

School of Information, University of Michigan, USA

Fraistat, Neil

nfraistat@gmail.com

Maryland Institute for Technology in the Humanities, University of Maryland, USA

Galloway, Patricia

galloway@ischool.utexas.edu

School of Information, University of Texas, Austin, USA

Kraus, Kari

karimkraus@gmail.com

College of Information Studies, University of Maryland, USA

Rehberger, Dean

Dean.Rehberger@matrix.msu.edu

MATRIX, Michigan State University, USA

Walter, Katherine

kwalter@unlnotes.unl.edu

Center for Digital Research in the Humanities, University of Nebraska, Lincoln, USA

Creative partnership between computer science and the humanities – what we now call "digital humanities" – is the cornerstone of the digital revolution. Cathy Davidson writes (2008) that "perhaps we need to see technology and the humanities not as a binary but as two sides of a necessarily interdependent, conjoined, and mutually constitutive set of intellectual, educational, social, political, and economic practices." Significant educational challenges exist, however, in creating a cadre of professionals who understand the intellectual context of digital humanities research and who are also capable of building the supporting infrastructure of digital collections, tools, and services (de Smedt 2002). The American Council of Learned Societies' groundbreaking report – *Our Cultural Commonwealth: Report of the Commission on Cyberinfrastructure for the Humanities and Social Sciences* – focuses attention on the need to "cultivate leadership in support of cyberinfrastructure from within the humanities

and social sciences, encourage digital scholarship, develop and maintain open standards and robust tools, and create extensive and reusable digital collections" (ACLS 2006, p. 4).

An international network of digital humanities centers creates and develops access to the digital documents, images, languages, sound, and film that constitute the human record and facilitate its understanding. In a quite separate but potentially symbiotic movement, graduate schools of information in the United States and elsewhere are producing technologically sophisticated professionals with deep backgrounds in and commitments to the humanities. Schools of Information, or "iSchools" have emerged from a two-decade long era of consolidation and reform, during which traditional schools of library science struggled with irrelevancy, diminished scale, and a fundamental societal transformation in the use of new and emerging technologies (Sawyer 2008). In North America, twenty-four iSchools have formed a caucus (<http://www.ischools.org/>) to advance a common agenda regarding the future of information studies. John Unsworth (2007) notes that digital humanities centers can establish new working relationships between humanities faculty and iSchool programs. iSchool faculty "are about half from other disciplines, and humanities computing is very much about information organization, ontologies, taxonomies, schema, preservation, interface design, and other issues that are studied and taught in [iSchool] programs. The [iSchool] connection also would help to activate the NEH/IMLS connection, as well as the NSF cyberinfrastructure connection."

While the move to develop digital humanities centers has demonstrated great successes, it has also meant the development of a number of unique but remote archives that are in danger of being lost. Universities in this digital age need to produce research and graduates that transcend traditional barriers and ways of working. The most influential origins of change wrought by information technology might well emerge from the humanities and information sciences, which consider most deeply the heritage and future of the human experience. The progress of this interdisciplinary field, however, requires new models of collaboration among the information sciences and the humanities disciplines. In this context it is worth noting that all of the iSchools involved in the digital humanities internship program (described below) have well established archives programs, from which they recruit graduate students to send to the participating DH Centers.

This paper for DH2010 presents a new model partnership initiative to help build curricular and

scholarly institutional infrastructures that leverage the existing and emerging capabilities of iSchools and digital humanities centers. With generous three-year support from the U.S. Institute of Museum and Library Services (IMLS), three iSchools and three digital humanities centers are placing graduate student interns for extended summer work experiences in digital humanities centers. The collaborators include the iSchools at the University of Maryland, the University of Texas and the University of Michigan; and the DH Centers at the University of Maryland (Maryland Institute for Technology in the Humanities), the University of Nebraska (Center for Digital Research in the Humanities), and Michigan State University (MATRIX). The partners are also developing a collaborative research program that draws on complementary areas of expertise and interest in the digital humanities and information studies. The project is in its second year, preparing to place a second set of interns in summer 2010, with a total of 18 internships offered over the duration of the project.

The DH2010 paper contextualizes the model internship program within the broader academic framework of the mission and activities of iSchools, including the humanities-oriented profiles of students, a curriculum that meaningfully combines--in holistic fashion--computational, legal, informational, cultural, social, and managerial content; and faculty research that crosses the two cultures of the humanities and sciences.

1. Student Profiles

The DH2010 paper presents a demographic analysis of the students enrolled at the three collaborating iSchools, demonstrating the affinity (and enriching the alliance) between iSchools and DH Centers. A majority of students enter into iSchool programs with undergraduate and/or graduate degrees in the arts and humanities, frequently outnumbering their more science-oriented peers by statistically significant margins. At Maryland's iSchool, for example, approximately 62 percent of current masters students (or 212 out of 343) obtained undergraduate degrees in English, History, Art History, Religion, Classics, and Philosophy. At Texas, well over half of the students have solid humanities backgrounds in literature, the arts, and especially in history; and show an impressive understanding of the values, styles, and methods of humanities researchers. At Michigan the humanities subject expertise of fully one-third of entering graduate students is integrated into a broader framework that incorporates techniques for systematically creating, managing, preserving and otherwise enhancing the value of cultural heritage information.

2. iSchool Curriculum

The DH2010 paper exposes how iSchools have implemented curricula of relevance to digital humanities centers, particularly in the area of cyberinfrastructure. At Michigan, for example, a suite of technology/systems-oriented courses teach students how to build and evaluate dynamic complex websites and databases; undertake preservation reformatting of books, graphical, and audiovisual resources; produce EAD finding aids and other access tools; and create and maintain online communities. At Maryland, students are exposed to the legal issues in managing information and the corpus of documents – such as donor agreements – that codify them. Copyright, privacy, freedom of information, and other topics pertinent to archives and digital libraries are also covered. At Texas, a series of courses in digitization for preservation and access is paired with courses in digital libraries and a sequence developing digital archiving practices to provide students with a range of skills and knowledge pertinent to preserving and providing access to humanities content.

3. Faculty Research

The DH2010 paper shows that iSchool faculty – an increasing number of whom have PhDs in arts and humanities disciplines – often conduct research with the potential to leverage and support the work of DH Centers. At Michigan, for example, Paul Resnick is pioneering work on recommender systems and the incentives that motivate eCommunities; Paul Conway is discovering how image quality issues in the large-scale digitization of cultural heritage resources impact innovative scholarship and use. At Texas, Gary Geisler is delving into improved interfaces for digital library presentations of materials from collections of the Harry Ransom Humanities Research Center; Patricia Galloway is investigating the use of tools from information retrieval and computational linguistics to frame digital corpora for the purposes of management and presentation. At Maryland, Derek Hansen and Kari Kraus recently received an NSF grant to understand and tailor alternate reality games for purposes of education and scholarly collaboration.

The DH2010 paper demonstrates how the first round of internships from the model program is establishing rich connections between iSchools and DH Centers. The model is sufficiently well articulated that it could further close working relationships between arts and humanities departments and DH Centers that wish to develop their own signature majors, minors, or concentrations in digital humanities. The paper concludes with a

summary of successes, progress, and road-blocks in implementing the new internship model.

4. Funding

This work was supported by The Institute of Museum and Library Services [grant number RE-05-08-0063-08].

References

- ACLS** (2006). *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. Washington, DC: American Council of Learned Societies. <http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf> (accessed 5 March 2010).
- Davidson, C.** (2008). 'Humanities 2.0: Promise, Perils, Predictions'. *PMLA*. **123(3)**: 707-717.
- de Smedt, K.** (2002). 'Some Reflections on Studies in Humanities Computing'. *Literary and Linguistic Computing*. **17(1)**: 89-101.
- Sawyer, S. and Rosenbaum, H.** (2008). 'I-Schools: Mice roaring or the future is now arriving?'. *2008 I-Conference*. Los Angeles, CA, 28 February – 2 March 2008. http://www.ischools.org/oc/conference08/ic08_PC_Papers.html.
- Unsworth, J.** (2007). 'Digital Humanities Centers as Cyberinfrastructure'. *Digital Humanities Centers Summit*. Washington, DC, 12 April. <http://www3.srl.illinois.edu/~unsworth/dhcs.html>.

Authorship Discontinuities of *El Ingenioso Hidalgo don Quijote de la Mancha* as detected by Mixture-of-Experts

Coufal, Christopher

coufalc@duq.edu

Duquesne University

Juola, Patrick

juola@mathcs.duq.edu

Duquesne University

In the literary world, authorship of great novels is like writing a great piece of music; while there may never be a perfect way to determine if someone wrote a particular work or not, equations and algorithms have been developed in information theory and statistics to help those trying to discover the true authorship of contested written works. Because no method is perfect, using a set of methods on the same works can be used to give high probabilities of authorship. The JGAAP system houses a collection of methods such as Histogram Distance and Manhattan Distance and event sets such as word bigrams and character trigrams to allow users to perform multiple tests on contested works to see if the supposed author is actually the author by comparing samples from both the contested author and other possible authors.

We apply this framework and show a notable discontinuity in the authorial style of the novel *El Ingenioso Hidalgo don Quijote de la Mancha*, better known as *Don Quijote* (or *Don Quixote*).

1. Background

While there have been skeptics and scholars alike that have doubted Miguel de Cervantes Saavade's true authorship of the entirety of *Don Quijote*, no one had tested whether or not Cervantes was in fact the true author of the whole of *Don Quijote*. The purpose of using the JGAAP system was to either give merit to or disprove this theory. By comparing to other authors who wrote works at about the same time *Don Quijote* was written, the JGAAP system would test to see if the text that Cervantes supposedly wrote was closer to the first volume of *Don Quijote* or closer to other authors of the same time period. If a definitive break could be established between where the program attributed Cervantes as the author and where it did not, that would suggest either a major style shift or the presence of another author different

from Cervantes, while no break at all would suggest that Cervantes was in fact the true author of the second volume of *Don Quijote*, assuming that he was also the author of the first volume.

2. Methods and Materials

For this authorship attribution, the program JGAAP 4.0 was downloaded from <http://www.jgaap.com>, developed by Patrick Juola at Duquesne University. The *Don Quijote* text used was acquired from Project Gutenberg at <http://www.projectgutenberg.org>. The full text of *Don Quijote* was then stripped of the introductions and separated into chapters by volume. The first volume was then set as the basis for Miguel de Cervantes' original authorship. Every third chapter, starting with chapter three, was used as the base case for Cervantes' work. Two other authors used for comparison, Francisco de Quevedo and Mateo Alemán, were also used. Quevedo's, *Historia de la vida del Buscón, llamado Don Pablos, ejemplo de vagamundos y espejo de tacaños* and Alemán's *Guzmán del Alfarache* were also taken from Project Gutenberg and broken into roughly the same number of chapter-type sections as the number of chapters used for Cervantes' *Don Quijote*. In order to make sure that Cervantes' was actually the author of volume one of *Don Quijote*, every chapter not used in the base case was compared to the base chapters, Quevedo's work, and Alemán's work. Each test used JGAAP's Normalize Whitespace, Strip Punctuation, and Unify Case canonizers on all of the documents. Five event sets - Word, WordBiGram, WordTriGram, WordTetraGram, and Word Length - were all paired with nine analysis methods - Camberra Distance, Cosine Distance, RN Cross Entropy, Histogram Distance, Kullback Leibler Divergence, Levenshtein Distance, Manhattan Distance, KS Distance, and Naive Bayes Classifier, for a total of 45 unique event set-analysis methods. Once the first volume of Cervantes' work was confirmed to be uniformly Cervantes', volume two of *Don Quijote* was tested in the same manner as the first volume in order to provide an accurate analysis.

We apply a mixture-of-experts approach to the evaluation of authorship. Each different method is treated as a single "expert" in different aspects of authorial style, and permitted to vote on who (among the candidates) is the author of any specific fragment. If all 45 test "experts" vote on Cervantes, we consider this to be strong evidence supporting his authorship, while if only 5 or so of the 45 consider Cervantes to be the most likely author, we consider this to be evidence *against*.

3. Results

As a result of the analysis on the second volume of *Don Quijote*, the JGAAP program indicated that starting at chapter 6 Cervantes was not the author. Out of the 45 tests run on each chapter, the chapters in the first volume had a mean of 37.54 occurrences of Cervantes as the author with a standard deviation of .852. The first five chapters of the second volume had a mean of 36.50 occurrences of Cervantes as the author with a standard deviation of 1.517. Chapters 6-74 of the second volume, however, had a mean of 4.90 occurrences of Cervantes as the author with a standard deviation of 1.436. This radical shift in authorship means either Cervantes completely shifted his writing technique or he did not write the latter 69 chapters of the second volume of *Don Quijote*.

4. Discussion

While there are people who are skeptic about the authorship of *Don Quijote*, nothing up until now has given those claims any grounds other than speculations based on inconsistencies in the text. Although this analysis does not guarantee that Cervantes did not write the last 69 chapters of the second volume, it does make the probability of that claim much greater. This, in part, is due to the fact that none of the tests in JGAAP has been tested enough to show that it will work for all documents. As further analysis of the methods continues, the results of the tests used in this authorship attribution will most likely validate these results. As tests and methods prove to not work, the analysis will be redone with these tests omitted from the analysis giving a more accurate result.

Entropy and Divergence in a Modern Fiction Corpus

Craig, Hugh

hugh.craig@newcastle.edu.au

School of Humanities and Social Science, University of Newcastle, Australia

The application of statistical methods to style is now well accepted in author attribution. It has found less favour in broader stylistic description. Louis Milic's pioneering quantitative work from the 1960s on the style of Jonathan Swift was vigorously contested by Stanley Fish, an attack which may well have had the effect of curbing enthusiasm for this kind of work. The other important exemplar is John Burrows' book on Jane Austen from 1987. I am not aware of any subsequent books of this kind.

In the proposed paper I aim to demonstrate the usefulness of two measures from Information Theory in the broad comparative analysis of text. One is entropy, which calculates the greatest possible compression of the information provided by a set of items considered as members of distinct classes (Rosso, Craig and Moscato). A large entropy value indicates that the items fall into a large number of classes, and thus must be represented by listing the counts of a large number of these classes. In an ecosystem, this would correspond to the presence of a large number of species each with relatively few members. The maximum entropy value occurs where each item represents a distinct class. Minimum entropy occurs where all items belong to a single class. In terms of language, word tokens are the items and word types the classes. A high-entropy text contains a large number of word types, many with a single token. A good example would be a technical manual for a complex machine which specifies numerous distinct small parts. A low-entropy text contains few word types, each with many occurrences, such as a legal document where terms are repeated in each clause to avoid ambiguity. Entropy is a measure of a sparse and diverse distribution versus a dense and concentrated one.

A second information-theory quantity which can serve for generalising about a set of texts is Jensen-Shannon Divergence (JSD). This gives a value to each set of items for the distance from a reference point, generally the mean for the whole grouping. This distance is calculated as the sum of divergences between the specimen and the mean for each of the classes represented in the set (Rosso, Craig and Moscato). In language terms the divergence value of a given text is the sum of the differences between the

counts for the text for each word type used in a corpus and the corpus mean count for that word type. Some texts use language in a way that closely corresponds to the norm of a larger set, others use some words more heavily, and others more lightly, than the run of a comparable corpus. JSD is a measure of normality in this specialised sense.

There are important caveats for interpreting these two measures of the properties of a text. Both are sensitive to text length, if for different reasons. Given a finite number of word types available to a given user of a given language, as a text sample grows, more of the pool is exhausted, and there is a greater tendency to recur to already-used word types. Thus in a novel the word tokens of a single sentence may well be all different word types, and have maximum entropy, but this is unlikely to be true of a paragraph, and still less so of a chapter. In the case of divergence from a mean, the law of averages means that for longer texts local idiosyncrasies tend to be balanced out by a larger body of less unusual writing and indeed by contrasting idiosyncrasies.

It is also important to rule out the idea that entropy and divergence values relate directly to quality. Entropy is related to a simpler measure, type-token ratio, sometimes called ‘vocabulary richness’. Yet ‘richness’ could scarcely be applied to a fighter plane manual, to revert to the example used above. One might associate divergence from the mean with originality or creativity, but it could just as well be the result of incompetence.

It is interesting that researchers have found genre to be a problem both with entropy work and with studies of intertextual distance when they are directed at authorship problems (Hoover, Labb   and Labb  ). From a different point of view, this sensitivity to genre is part of what makes the methods valuable for a more general assessment of the style of texts.

The corpus for the study in the paper consists of 377 fiction texts, being the first 25,000 words of all the texts with 25,000 words or more in the British National Corpus ‘Imaginative Fiction’ section. This amounts to 15,421,915 words in all. The texts are predominantly prose fiction published in the United Kingdom in the 1990s, taken from a wide variety of sources, short stories as well as novels, intended for young and young adult audiences as well as for a general readership. The usefulness of JSD results depend on the validity of the point of reference chosen. In the present study the mean of this large collection of texts of very varied authorship and genre, within the larger text type ‘imaginative fiction’, should be a good approximation of the mean for contemporary fiction in general.

At the time of writing this proposal work on this corpus with these methods is at an early stage, but

there are some preliminary findings. The first is that entropy and divergence are positively correlated in this corpus. As density decreases and a wider range of word types are used for the same extent of text, samples diverge more from the mean. The individual exceptions to these broad tendencies are instructive and some individual examples will be discussed.

It is also possible to see at this early stage that within the universe of prose fiction these two quantities align with more impressionist views of style. High entropy fiction texts follow a traditional ‘high style’. Their progression is linear, continuing to move on to new vocabulary, while low entropy texts retrace their steps and return to already used words. High entropy texts are demanding of the reader and dense in information. They constantly move to new mental territories; they are taxing and impressive. Low entropy texts are reassuring and familiar. They are implicit in their signification, assuming common knowledge, while high-entropy texts specify and create contexts for themselves. High-entropy texts contain more description and narrative, while low-entropy texts contain more dialogue.

The challenge for computational approaches to style is to use the power of statistics working on the abundant data available from texts to reveal tendencies which are important, yet would otherwise be invisible, or remain in the realm of the impressionistic. The argument of the proposed paper is that the entropy and divergence of words provide two useful ways of understanding fundamental properties of texts. Entropy and divergence are soundly based in statistical theory and informative on two fronts. They open the way to density and normality as fundamental ways of thinking about style; and they serve to place particular texts in relation to sets of comparison texts and thus to map them in a conceptual space. Short stories and novels may be virtual worlds, intensely personal meditations, and human dramas of love and conflict, but they are also sets of vocabulary items used with a given frequency, and it is surprising how much an analysis of that base level of their existence can reveal about them.

References

- British National Corpus, version 2** (2001). *BNC World*. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.
- Burrows, J. F.** (1987). *Computation into Criticism: A Study of Jane Austen and an Experiment in Method*. Oxford: Clarendon.
- Fish, Stanley** (1980). ‘What Is Stylistics and Why Are They Saying Such Terrible Things About It?’. *Is*

There a Text in This Class?. Cambridge MA: Harvard University Press, pp. 68-96.

Hoover, David (2003). 'Another Perspective on Vocabulary Richness'. *Computers and the Humanities*. **37**: 151-78.

Labbé, Cyril, Labbé, Dominique (2006). 'A Tool for Literary Studies: Intertextual Distance And Tree Classification'. *Literary and Linguistic Computing*. **21.3**: 311-26.

Milic, Louis T. (1967). *A Quantitative Approach to the Style of Jonathan Swift*. Mouton: The Hague.

Rosso, Osvaldo, Craig, Hugh, Moscato, Pablo (2009). 'Shakespeare and Other English Renaissance Authors as Characterized by Information Theory Complexity Quantifiers'. *Physica A*. **388**: 916-26.

Objective Detection of Plautus' Rules by Computer Support

Deufert, Marcus

mdeufert@equa.net

Department of Classics, University of Leipzig,
Germany

Blumenstein, Judith

jblumenstein@equa.net

Department of Classics, University of Leipzig,
Germany

Trebesius, Andreas

atrebesius@equa.net

Department of Classics, University of Leipzig,
Germany

Beyer, Stefan

sbeyer@equa.net

Natural Language Processing Group, Institute of
Mathematics and Computer Science, University of
Leipzig, Germany

Büchler, Marco

mbuechler@equa.net

Natural Language Processing Group, Institute of
Mathematics and Computer Science, University of
Leipzig, Germany

The metre of the Roman comic poet Plautus (flourished ca. 200 B.C.) still leaves one mystified. Although the scientific work of the 19th and early 20th century has established a number of important rules and licences, the exact range of these laws and licences remains a matter of debate. Taking into account these many open questions it is not surprising that metrical studies (as well as the important editions) of Plautus still display a huge amount of discrepancy in their handling of Plautine metre. The specific problem consists of the large number of transmitted verses in the Plautine corpus and the great complexity and diversity of competing explanations of remarkable metrical phenomena.

Therefore, until now the results of scholarship often fail to convince, since they are based either on a limited textual basis or deal with a specific metrical phenomenon from the perspective of a single law or licence without taking into account competing explanations.

This paper will cover a wide range of previous research from both Classics (Lotman 2000) and literature (Garzonio, 2006) to several techniques in the field of Computer Science (Heyer et al., 2008

and Volk, 2007). Metric analyses can be already be computed on German poems with only a small set of rules (Bobhausen, 2009). Results imply, however, that foreign words are especially difficult to handle. In contrast to this, ancient texts pose a different problem as lots of variations of an original often exist. For this reason metric analysis can be divided into three different tasks:

- **Task 1:** Dealing with different **variations** and variations of variations (Andreev, 2009 and Rehbein, 2009). Within this paper, a primary version of a verse is defined by researchers from Classics. Differences of variants in relation to the primary version are highlighted as described in (Büchler et al., 2009¹ and Rehbein, 2009). The variance caused by transmissions of several authors is also important to consider, however, when working with fuzziness. Consequently a set of possible metric analysis annotations are suggested rather than just one result.
- **Task 2:** **Applying a metric rule-set** to a text corpus (Bobhausen, 2009 and Fusi 2008). Within this research - similar to part of speech tagging (Heyer et al., 2003) – a set of rules is applied to text. However, only the most probable metric candidate is selected. In contrast to that research (Bobhausen, 2009 and Fusi, 2008), the approach in this paper scores several possible metric analyses.
- **Task 3:** **Training of a metric ruleset** based on manually annotated data from researchers. Typically, a fixed set of rules is taken as presumed, however, new rules need to be added manually. This paper thus also focuses on the computation of new rules. The importance of this step is motivated by the Theory of Selective Perception. Based on this, new and uncommon rules are determined by a computer model that is both objective and independent rather than selective like a human being.

In the field of natural language processing the task of tagging text is quite similar to part of speech tagging (POS). Typically, for such a tagger a Hidden Markov Model (Heyer et al., 2008) is trained and is traversed by dedicated algorithms like the Viterbi algorithm (Heyer et al., 2008). However, the already mentioned fuzziness of text variants makes both the training and traversing steps difficult. Furthermore, in the training step it is necessary to observe data on a larger window than the typical memory of 2 or 3. This would increase the complexity drastically during the trainings phase. Within the applying phase the Viterbi algorithm is typically used (Heyer et al., 2003). This algorithm reduces all paths locally except the most probable one. In metric analysis however this assumption is quite critical since due to syllable

fusion an senarius is not required to have 12 but can also consist of 17 or 18 syllables.

Motivated by the aforementioned problems of existing approaches this paper describes a three step approach. In a first step possible syllables are computed. This is simply done by using training data. In contrast to German poems (Bobhausen, 2009) the approach is aware of possible fusions of syllables. In the second step all possible combinations are computed instantly removing candidates that do not fulfil the metric requirements. The training itself is done by distance-based co-occurrences (Büchler, 2008) on metric tags. In the last step metric candidates are scored based on both the training data as well as the variance of the alternatively transmitted variances. All relevant candidates are selected by researchers of Classics. The remaining metric analysis is represented in a dedicated visualisation highlighting the differences of several variants to the primary version (Büchler, 2009 and Rehbein, 2009).

As an outcome of this paper several results will be shown. Besides the difference visualisation both results and experiences in training and application of a metric model will be provided.

Both the expected results and the developed software, which can be easily adapted to other ancient poets, will give an original input to the research community and motivate and enable further investigations in the same spirit.

References

- Andreev, V. S.** (2009). 'Patterns in Style Evolution of Poets'. *Digital Humanities 2009*. Pp. 52-53.
- Bobhausen, K.** (2009). 'Automatisches Metrisches Markup'. *Digital Humanities 2009*. Pp. 69-72).
- Büchler, M.** (2008). *Elemente einer Forensischen Linguistik* Working report.
- Büchler, M., Geßner, A.** (2009). 'Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts'. *2009 Chicago Colloquium on Digital Humanities and Computer Science*. Chicago, Nov. 2009.
- Fortson, B. W.** (2008). 'IV, Language and Rhythm in Plautus'. *Synchronic and Diachronic Studies*. Berlin / New York.
- Fusi, D.** (2009). *An Expert System for the Classical Languages: Metrical Analysis Components*. <http://www.fusisoft.it/Doc/ActaVenezia.pdf> <http://www.fusisoft.it/Chiron/Metrics/Default.aspx> (accessed Nov. 10th 2009).

Garzonio, S. (2006). 'Italian and Russian Verse: Two Cultures and Two Mentalities'. *Studi Slavistici*. III: 187-198.

Heyer, G., Quasthoff, U., Wittig, T. (2008). *Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse. 2nd edition*. W3L-Verlag.

Lotman, M.-K. (2009). 'Word-ends and Metrical Boundaries in Ancient Iambic Trimeter of Comedy'. *Studia Humaniora Tartuensa*. 1: 1-16. <http://www.ut.ee/klassik/sht/2000/lotman1.pdf> (accessed Nov., 10th 2009).

Questa, C. (2007). *La metrica di Plauto e di Terenzio*. Urbino.

Rehbein, M. (2009). 'Multi-Level Variation'. *Digital Humanities Conference Abstracts*. 2009 2009, pp. 11-12.

Volk, A. (2007). 'Rhythmic Similarity based on Inner Metric Analysis'. *Utrecht Summer School Multimedia Retrieval*. Utrecht, Aug. 2007.

Notes

1. In this paper the same visualisation is used to highlight differences of "literal" citations in a historic context.

The ecology of longevity: the relevance of evolutionary theory for digital preservation

Doorn, Peter

peter.doorn@dans.knaw.nl

Data Archiving and Networked Services,
Netherlands

Roorda, Dirk

dirk.roorda@dans.knaw.nl

Data Archiving and Networked Services,
Netherlands

Software and data can be considered as digital organisms that function in a "digital ecosystem" of computers. The concept of ecology has been borrowed from biology by other disciplines for explaining or describing a variety of phenomena. In some cases ecology and other concepts from evolutionary theory are only used as metaphors; in other cases attempts have been made to apply an adapted version of the theory to the evolution of non-biological phenomena.

We think it makes sense to borrow the notions from evolutionary theory in thinking about digital longevity. In this paper we will explore the potential of Darwin's theory as an explanatory framework for digital survival.

The construction of a theoretical foundation serves to answer questions of why some criteria or characteristics will guarantee the "survival" of digital objects better than other ones. Taking an evolutionary view will also make clear that there is no such thing as digital permanence for eternity: some objects only have better *chances* to survive than other ones. In computing technology, there is a struggle of survival of the fittest going on. In this struggle, new technologies arise as modifications or adaptations of earlier technologies and the older ones die out when newer technologies are stronger or better suited for their tasks. The digital objects that are already in existence have to be adapted to the new technological surroundings, otherwise they become extinct.

Spencer originally coined the phrase "survival of the fittest" in 1864, drawing parallels between his ideas of economics and Darwin's theory of evolution, which is driven by "natural selection". With respect to digital objects, it is people who are actively or passively involved in making the selections, thus deciding which digital objects survive and which do not.

As electronic digital data can only be understood using computers and software to "translate" them into visible or audible form, the media and data formats that are specific for hard- and software change according to the same evolutionary principles. If we accept this view, we can start to ask ourselves: which characteristics (comparable to the "genetic properties" of living organisms) may influence the chances of digital survival of data? This is however not unproblematic, as it is typically with hindsight that we see which (traits in) a biological species have survived and how the evolutionary process took place. The explanatory power of evolution theory is *a posteriori*, not *a priori*. It may therefore be difficult to predict which traits are good for survival.

We will explore the possible use of the concept of evolvability, which is usually defined as the ability of a population of organisms to generate genetic diversity, hence giving a measure of an organism's ability to evolve. Maybe there is a parallel here with respect to digital objects. For instance, if we look at several formats for Microsoft Word (.doc, .rtf, .html, .xml (2003), .ooxml) then we see an increase in usability/interchangeability, and hence probably: evolvability.

We can break down the survival problem to questions concerning:

- The physical attributes of the media (tape, disk, etc.)
- The media format (density, size, etc.)
- The data content (integrity of the bits and bytes)
- The data format (the structure of the bits and bytes)
- The metadata content (the substantial description of the data)
- The metadata format (the format in which the metadata is described)
- The interlinking (the degree to which data is linked both internally and externally); a web of interlinked information is an ecosystem of its own.

We will demonstrate how digital preservation strategies such as technology preservation, software emulation, and data migration fit in an overarching evolutionary framework. The ecological approach also shows that it makes no sense to try to express the time horizon for the preservation of digital objects as a specific or indefinite period of time, but that we can better think in terms of "chances of survival".

The evolutionary framework can be used to argue why certain attributes and formats are more likely to survive than others. Also, analogous to natural selection, we will make clear that there

is no single "best" strategy for survival of digital data. Some factors simply increase the chances of digital longevity, whereas other factors reduce these chances. Good factors for longevity may be bad for other desired characteristics. For example: stripping executable information from data improves its longevity, but hinders its functionality. It may also be so that some factors are intensely ambiguous for longevity. We may now think that "wrapping" text in WordPerfect in the 1990s was (with hindsight) not so good for survival, and that packaging it in Microsoft Word seems acceptable. This is probably related to the status (or market dominance) of the software packages. Similarly, packaging data in SGML in 1990 might have been not so good, while packaging it in XML in 2009 seems excellent. In the end, the environment determines what was good and what was bad for longevity.

So, when the whole "technological ecosystem" changes, what was well adapted before the change may appear to be ill suited in the next technological phase. Digital preservation strategies can use the principle of digital selection in order to maximize the adaptation of digital objects to their environment, thus increasing their chances of digital longevity.

Whether it makes sense to apply evolution theory to digital curation can be studied by looking at a number of parallels in other scientific domains. We will deal briefly with attempts to use Darwin's ideas in the social sciences and in technology. In the social sciences the idea of a "social ecology" was already applied and empirically tested in the 1920s by, among others, Robert Park and Ernest Burgess of the "Chicago School" of urban ecology. With respect to man-created systems it is probably better to use the ideas on evolution by Lamarck. Lamarckism is the idea that an organism can pass on characteristics that it acquired during its lifetime to its offspring (also known as heritability of acquired characteristics or soft inheritance).

Several researchers have proposed that Lamarckian evolution may be accurately applied to cultural evolution. Human culture can be looked upon as an ecological niche-like phenomenon, where the effects of cultural niche construction are transmissible from one generation to the next. Ecological notions on the evolution of software, in which ideas and characteristics of programming languages compete with each other, have been formulated in information science. Inheritance is an important concept with an evolutionary basis.

The development of open source software has also been described as evolving in a Lamarckian fashion. Ensuring free access and enabling modification at each stage in the process means that the evolution of software occurs in the fast Lamarckian mode: each

favourable acquired characteristic of others' work can be directly inherited.

Kauffman and Dennett point out the parallels between biological evolution and technological evolution. They distinguish two stages: (i) explosion of the number of greatly different designs when there are still many unoccupied niches; (ii) microevolution where the existing designs are optimised for competition in existing niches.

It is also useful to compare the selection and survival of digital information with that of analogue information. In both cases there is "information selection" and evolution. What makes the digital world so different from the analogue world?

Next we will treat a few examples of the evolution of computing technologies, software, file formats and data sets, which will illustrate how well evolutionary theory is suited for explaining what has happened empirically. It makes sense to look backwards and use evidence from the – still very short – historical evolution of computing technology since the 1940s and '50s. Can we still read the first image, the first word processor file, the first database, the first web page, email or pdf-file? If yes, how come this is still possible? If not, why? And how did the formats of those data types (probably equivalent to the taxonomical rank of the *genus* in biology) evolve, and which abandoned file formats ("species") can still be read today?

Maybe there is a manifestation of genotype/phenotype here, where the application can be considered as the phenotype. Applications struggle with each other in the "econosphere" of consumers. The surviving application dictates the data format. If there are two strong surviving applications in a domain, you get peer-to-peer data convertors. If there are many, weaker survivors, you get interchange formats, which are better for preservation. A familiar lesson is not to rely on monopolists.

Finally, there is a marked tendency among data curators to set criteria for "trusted digital repositories", the nature reserves of the digital world. On the basis of the evolutionary ideas expressed in this paper, it makes sense not to make such criteria too narrow. They should be chosen in such a way that they make use of the "natural" evolutionary processes of technology and digital objects, making sure that what is threatened by extinction can be rescued in an effective way.

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

Shippensburg University, USA

Sachs, Jon

jsachs@alcor.concordia.ca

Concordia University, Canada

This paper will report on the first year of a 3-year grant funded by the *Fonds québécois 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 nested 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.

Does Size Matter? Authorship Attribution, Small Samples, Big Problem

Eder, Maciej

maciej_eder@poczta.onet.pl

Pedagogical University, Krakow, Poland

The aim of this study is to find a minimal size of text samples for authorship attribution that would provide stable results independent of random noise. A few controlled tests for different sample lengths, languages and genres are discussed and compared. Although I focus on Delta methodology, the results are valid for many other multidimensional methods relying on word frequencies and "nearest neighbor" classifications.

In the field of stylometry, and especially in authorship attribution, the reliability of the obtained results becomes even more essential than the results themselves: failed attribution is much better than false attribution (cf. Love, 2002). However, while dozens of outstanding papers deal with increasing the effectiveness of current stylometric methods, the problem of their reliability remains somehow underestimated. Especially, the simple yet fundamental question of the shortest acceptable sample length for reliable attribution has not been discussed convincingly.

In many attribution studies based on short samples, despite their well-established hypotheses, convincing choice of style-markers, advanced statistics applied and brilliant results presented, one cannot avoid a very simple yet uneasy question: whether those impressive results could be obtained *by chance*, or at least positively affected by *randomness*? This question can be also formulated in a different way: if a cross-checking experiment with numerous short samples were available, would the results be just as satisfying?

1. Hypothesis

It is commonly known that word frequencies in a corpus are random variables; the same can be said about any written authorial text, like a novel or poem. Being a probabilistic phenomenon, word frequency strongly depends on the size of the population (i.e. the size of the text used in the study). Now, if the observed frequency of a single word exhibits too much variation for establishing an index of vocabulary richness resistant to sample length (cf.

Tweedie and Baayen, 1998), a multidimensional approach – based on several probabilistic word frequencies – should be even more questionable.

On theoretical grounds, we can intuitively assume that the smallest acceptable sample length would be hundreds rather than dozens of words. Next, we can expect that, in a series of controlled authorship experiments with longer and longer samples tested, the probability of attribution success would at first increase very quickly, indicating a strong correlation with the current text size; but then, above a certain value, further increase of input sample size would not affect the effectiveness of the attribution. In any attempt to find this critical point in terms of statistical investigation, one should be aware, however, that this point might depend – to some extent – on the language, genre, or even the text analyzed.

2. Experiment I: Words

A few corpora of known authorship were prepared for different languages and genres: for English, Polish, German, Hungarian, and French novels, for English epic poetry, Latin poetry (Ancient and Modern), Latin prose (non-fiction), and for Ancient Greek epic poetry; each contained a similar number of texts to be attributed. The research procedure was as follows. For each text in a given corpus, 500 randomly chosen single words were concatenated into a new sample. These new samples were analyzed using the classical Delta method as developed by Burrows (2002); the percentage of attributive success was regarded as a measure of effectiveness of the current sample length. The same steps of excerpting new samples from the original texts, followed by the stage of "guessing" the correct authors, were repeated for the length of 600, 700, 800, ..., 20000 words per sample.

The results for a corpus of 63 English novels are shown on Fig. 1. The observed scores (black points on the graph; grey points will be discussed below) clearly indicate the existence of a trend (solid line): the curve, climbing up very quickly, tends to stabilize at a certain point, which indicates the minimal sample size for the best attributing rate. It becomes quite obvious that samples shorter than 5000 words provide a poor "guessing", because they can be immensely affected by random noise. Below the size of 3000 words, the obtained results are simply disastrous. Other analyzed corpora showed that the critical point of attributive success could be found between 5000 and 10000 words per sample (and there was no significant difference between inflected and non-inflected languages). Better scores were obtained for the two poetic corpora: English and Latin (3500 words per sample were enough for good results), and, surprisingly, the corpus of Latin prose

(its minimal effective sample size was of some 2500 words; cf. Fig. 2, black points).

3. Experiment II: Passages

The way of preparing samples by extracting a mass of single words from the original texts seems to be an obvious solution for the problem of statistical representativeness. In most attribution studies, however, shorter or longer *passages* of disputed works are usually analyzed (either randomly chosen from the entire text, or simply truncated to the desired size). The purpose of the current experiment was to test the attribution effectiveness of this typical sampling. The whole procedure was repeated step by step as in the previous test, but now, instead of collecting individual words, sequences of 500 words (then 600, 700, ..., 20000) were excerpted randomly from the original texts.

Three main observations could be made here: 1. For each corpus analyzed, the effectiveness of such samples (excerpted passages) was *always* worse than the scores described in the former experiment, relying on the "bag-of-words" type of sample (cf. Fig. 1 and 2, grey points). 2. The more inflected the language, the smaller the difference in correct attribution between both types of samples, the "passages" and the "words": the greatest in the English novels (cf. Fig. 1, grey points vs. black), the smallest in the Hungarian corpus. 3. For "passages", the dispersion of the observed scores was *always* wider than for "words", indicating the possible significance of the influence of random noise. This effect might be due to the obvious differences in word distribution between narrative and dialogue parts in novels (cf. Hoover, 2001); however, the same effect was equally strong for poetry (Latin and English) and non-literary prose (Latin).

4. Experiment III: Chunks

At times we encounter an attribution problem where extant works by a disputed author are doubtless too short for being analyzed in separate samples. The question is, then, if a concatenated *collection* of short poems, epigrams, sonnets, etc. in one sample (cf. Eder and Rybicki, 2009) would reach the effectiveness comparable to that presented above? And, if concatenated samples are suitable for attribution tests, do we need to worry about the size of the original texts constituting the joint sample?

The third experiment, then, was designed as follows. In 12 iterations, several word-chunks were randomly selected from each text into 8192-word samples: 4096 bi-grams, 2048 tetra-grams, 1024 chunks of 8 words in length, 512 of 16 words, and so on, up to 2 chunks of 4096 words. Thus, all the samples in

question were 8192 words long. The obtained results were very similar for all the languages and genres tested. As shown in Fig. 3 (for the corpus of Polish novels), the effectiveness of "guessing" depends to some extent on the word-chunk size used. Although the attributive scores are slightly worse for long chunks within a sample (4096 words or so) than for bi-grams, 4-word chunks etc., every chunk size could be acceptable to constitute a concatenated sample.

However, although this seems to be an optimistic result, we should remember that this test would not be feasible on really short poems. Epigrams, sonnets etc. are often masterpieces of concise language, with a domination of verbs over adjectives and so on, and with a strong tendency to compression of content. For that reason, further investigation is needed here.

5. Conclusions

The scores presented in this study, as obtained with classical Delta procedure, would be slightly better when solved with Delta Prime, and worse if either Cluster Analysis or Multidimensional Scaling is used (a few tests have been done). However, the shape of all the curves, as well as the point where the attributive success rate becomes stable, are quite identical for each of these methods. The same refers to different combinations of style-markers' settings, like "culling", the number of the Most Frequent Words analyzed, deleting/non-deleting pronouns, etc. – although different settings provide different "guessing" (up to 100% for the most efficient), they never affect the shape of the curves. Thus, since the obtained results are method-independent, this leads us to a conclusion about the smallest acceptable sample size for future attribution experiments and other investigations in the field of stylometry. It also means that some of the recent attribution studies should be at least re-considered. Until we develop style-markers more precise than word frequencies, we should be aware of some limits in our current approaches. As I tried to show, using 2500-word

samples will hardly provide a reliable result, to say nothing of shorter texts.

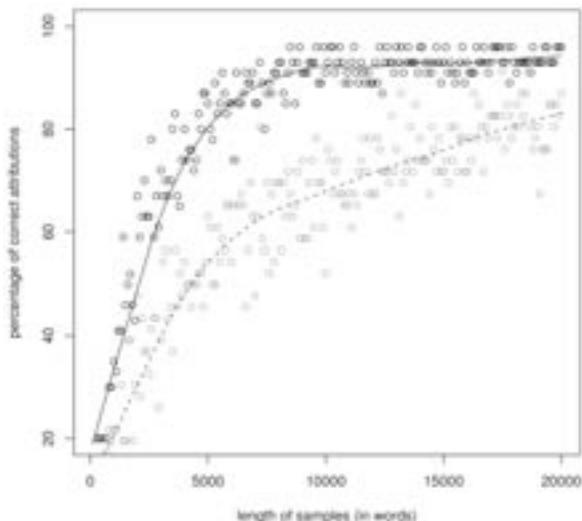


Figure 1: English novels

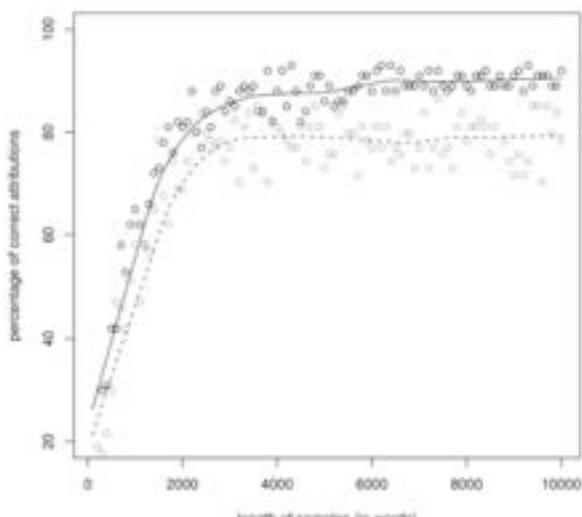


Figure 2: Latin prose

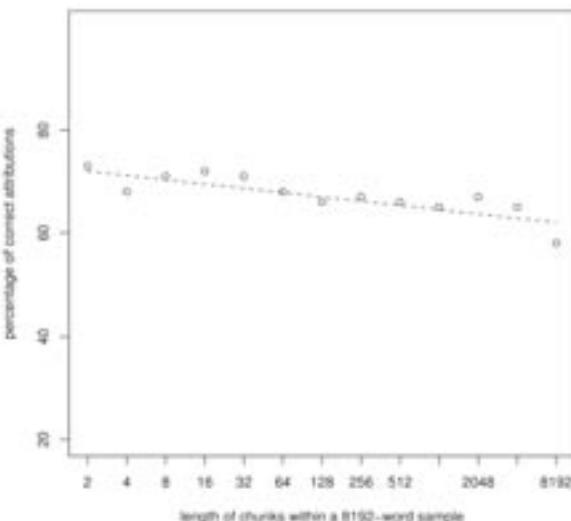


Figure 3: Polish novels

References

- Burrows, J. F.** (2002). 'Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship'. *Literary and Linguistic Computing*. **17**: 267-287.
- Craig, H.** (2004). 'Stylistic Analysis and Authorship Studies'. *A Companion to Digital Humanities*. S. Schreibman, R. Siemens and J. Unsworth (ed.). Blackwell Publishing, pp. 273-288.
- Eder, M., Rybicki, J.** (2009). 'PCA, Delta, JGAAP and Polish Poetry of the 16th and the 17th Centuries: Who Wrote the Dirty Stuff?'. *Digital Humanities 2009: Conference Abstracts*. University of Maryland, College Park, pp. 242-244.
- Hoover, D. L.** (2001). 'Statistical Stylistic and Authorship Attribution: an Empirical Investigation'. *Literary and Linguistic Computing*. **16**: 421-444.
- Hoover, D. L.** (2003). 'Multivariate Analysis and the Study of Style Variation'. *Literary and Linguistic Computing*. **18**: 341-360.
- Juola, P., Baayen R. H.** (2005). 'A Controlled-corpus Experiment in Authorship Identification by Cross-entropy'. *Literary and Linguistic Computing*. **Suppl. Issue 20**: 59-67.
- Love, H.** (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.
- Rudman, J.** (1998). 'The State of Authorship Attribution Studies: Some Problems and Solutions'. *Computers and the Humanities*. **31**: 351-365.
- Rybicki, J.** (2008). 'Does Size Matter? A Re-examination of a Time-proven Method'. *Digital*

Humanities 2008: Book of Abstracts. University of Oulu, pp. 184.

Tweedie, J. F. and Baayen, R. H. (1998). 'How Variable May a Constant be? Measures of Lexical Richness in Perspective'. *Computers and the Humanities*. **32**: 323-352.

Finding Stories in the Archive through Paragraph Alignment

Esteva, Maria

maria@tacc.utexas.edu

Texas Advanced Computing Center (TACC),
University of Texas at Austin, USA

Xu, Weijia

xwj@tacc.utexas.edu

Texas Advanced Computing Center (TACC),
University of Texas at Austin, USA

We present research showing the possibility of finding stories in a digital text archive through computational methods. Referring to the concept of "archival bond", we define stories as formed by documents that relate to a target activity. We developed a method called *paragraph alignment* to find such documents and an interactive visualization to discover connected stories in context with provenance.

Our method was applied to the challenges presented by the digital archive of a multinational philanthropic organization who awarded grants to cultural, scientific, and social welfare activities (1985-2005). Over fifteen years, the staff members deposited their work documents in individual directories on a shared server without following any record-keeping rule. These documents reflect the organization's activities in the areas of Science and Education, Art and Humanities, and Social Welfare. They also reflect the staff members' records creation practices, afforded by the cut and paste function of the word processor and the possibility to collaborate through the network. These digital aggregations are sometimes perceived as chaotic, defined as ROT (redundant, outdated and trivial,) and deemed disposable (Henry, 2003; AIIM, 2009; Public Records Office, 2000). Yet they are ubiquitous in the networked servers of many organizations, so our goal was to find a method to make sense of the text records within.

1. Archival Bond

A fundamental concept in archival theory, known as archival bond, describes the relationships between documents in an archive as essential properties of the documents (Duranti, 1997). While all the documents in a collection are bonded through the collection's structure (McNeil, 2000), there are stronger relationships between sub-groups of documents that belong to the same function and/

or activity. In the case of disorganized electronic text archives in which the structure is nonexistent or loose, we suggest that the relationships among documents be defined based on their content referring to a target activity. By finding trails of documents that narrate stories about activities in context with provenance, we aim to establish order, identify structure, and learn about the archive's creators.

2. Paragraph Alignment (PA)

We observed that in our archive, similar paragraphs about an activity are repeated across short - memos and press releases - and long documents - annual reports and board meeting minutes. As a group, these documents tell the story of an activity. We also observed that in many documents the same personal names, places, and institutions are mentioned in relation to different activities, and that documents that use similar terms may not be associated with the same activity. The traditional cosine similarity method measures global similarity between documents. Given the characteristics noted in this archive, we considered that calculating global similarity was not efficient to identify all the documents about a target activity. Instead, we draw from local alignment, a method used in bioinformatics to evaluate local similarity between sequences (Gusfield, 1997).

While biological sequences evolve throughout time owing to constant mutation events, the parts of the sequences that directly participate in cellular activities remain relatively stable. Therefore, global similarity between two sequences is often less important than the local similarity, which is defined by the highest similarity between any two substrings from two sequences. Efficient methods for computing sequence similarities often follow a framework in which sequences are broken into n-gram for similarity computations and then assembled to derive an overall similarity (Wu et al., 1990). Here we adapt a similar approach that we call paragraph alignment to determine archival bond between documents.

Our method contrasts with previous work on document segmentation (Hearst, 1994). Rather than measuring inter-paragraph similarity within one document to identify subtopic structure, our approach focuses on comparing the similarity between document segments to identify topics across a collection. Hence the primary goal of document segmentation is to minimize the variation of length between documents for subsequent similarity comparison.

3. Methodology

Figure 1 shows the workflow of our approach.

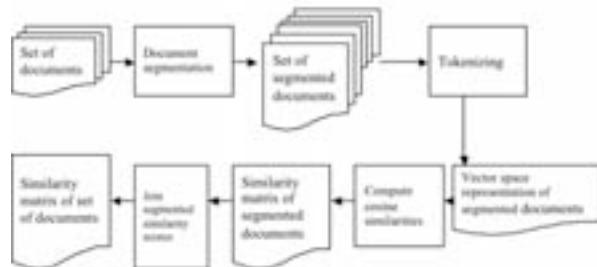


Figure 1

Each document in a set is broken into one or more ordered segments based on the paragraphs in the document. If the length of a segment (including spaces) is less than a pre-defined minimum number of characters threshold (MNCT), the segment is merged with the following segment. We used MNCT of 1000, 750, and 500 characters. For each set of document segments we create a matrix of TFIDF weighted term frequencies after stop-words removal (McCallum, 1996), and then calculate the cosine similarity between every other segment (Salton, 1988). We then process the resultant matrix to derive similarity scores between document pairs, which are defined as the maximum similarity score between their segments. For evaluation, we compare the results of the different MNCT with those obtained by calculating cosine similarity as a measure of global similarity between the documents.

We tested the method in a set of 714 documents from the year 1997 with eight authors. Date and authorship were preserved in the documents' file name. The evaluation was based on assessing seven document test-groups. A team member familiar with the archive selected five query documents, each corresponding to a different activity (test-groups 1, 2, 4, 5, 6) and two containing summaries of various activities (test-groups 3 and 7). For each query document, the team member also identified a set of related documents. For each test-group, both the cosine similarity and the paragraph alignment methods returned a list of documents ranked from more similar to less similar. The team member checked the results against the content of the corresponding document labeling the ranked document as a "true positive" if it was related to the query document; otherwise the document was labeled as "false positive". Results were checked until the last true positive was found.

4. Results

Test-group	1	2	3	4	5	6	7
Number of true positives	21	5	9	17	19	19	43
Number of false positives	28	36	6	205	88	53	103
Cosine	10	49	10	20	87	46	47
PA 1000	6	27	12	6	34	38	70
PA 750	7	186	11	17	100	107	71
PA 500	1	1	1	1	1	2	1
PA 100	2	2	12	2	4	1	9
Number of document segments based on number of characters	1	1	9	1	3	1	7
PA 1000	1	1	6	1	2	1	6
PA 750	1	1	1	1	1	1	1
PA 500	1	1	1	1	1	1	1
PA 100	1	1	1	1	1	1	1

Table1

The results show that the PA method with a MNCT of 750 characters returned better results five out of seven times (test-groups 1, 2, 4, 5, 6 and 7). For test-group 7, the best results were obtained with a MNCT of 500 characters. In this case the query document contained summaries of five different projects accomplished during 1997, each mentioned in other documents in the set. This suggests that although related documents in the set may not share similar global word distributions, they share similar word distributions in some of their segments. While the efficiency of the different MNCT depends on the particular word distribution of the documents that are being compared, in general, the smaller the MNCT used the higher the documents with less global similarity are ranked by the PA method. The PA method did not work for test-group 3 which contains sentences about activities most of which are not mentioned in other documents in the set. Figure 2 shows a plot of the results of test-group 1 in which the PA method with a MNCT of 750 characters performed the best.

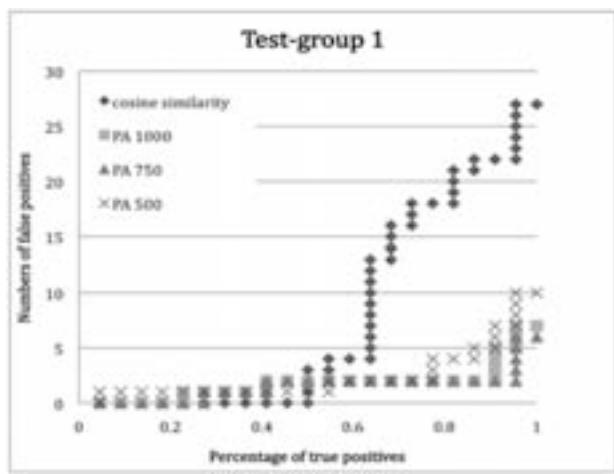


Figure 2

The test-group one (Figure 2) contains documents about a program to train young orchestra directors. The query document is a memo including a brief description of the project and estimated costs for lodging and travel. Returned true positive

documents of five authors include: other planning documentation, correspondence with potential contributors, the call for applications, a press release, a list of participants, the musical program, and various reports.

5. Visualization

Through an interactive visualization (Figure 3) users can follow the connections between documents to identify stories (PREFUSE). Each document is labeled with a color corresponding to its author. The connectivity between the documents allows the identification of a) stories, b) the authors involved in a given activity, and c) connected stories. As the user interacts with the visualization, the structure of the archive takes shape. Below is a snapshot of the visualization interface showing stronger connections between a group of documents.



Figure 3

6. Conclusions

This research has implications for the retention of digital archives. Using the concept of archival bond as a framework we aim to make sense of ROT archives and to unveil their stories. The results show that for documents that share similar paragraphs, local similarity matters to identify an archival bond. The same characteristic is observed in the biological sequence analysis that inspired our method.

7. Acknowledgments

This work was supported through a National Archives and Records Administration (NARA) supplement to the National Science Foundation Cooperative Agreement (NSF) TERAGRID: Resource Partners, OCI-0504077.

References

- Henry, Linda J. (2003). 'Appraisal of Electronic Records'. *Thirty Years of Electronic Records*. B. I. Ambacher (ed.). Maryland: The Scarecrow Press, pp. 38.

'Best Practices for Information Organization and Access'. *AIIM*. 2009 <http://www.aiim.org/infonomics/best-practices-for-IOA.aspx> (accessed 29 October 2009).

'Guidance for an Inventory of Electronic Records: a Toolkit'. *Public Records Office*. 2000 http://www.nationalarchives.gov.uk/documents/inventory_toolkit.pdf (accessed 29 October 2009).

Duranti, L. (1997). 'The Archival Bond'. *Archives and Museum Informatics*. 213.

McNeil, H. (2000). *Trusting Records: Legal, Historical and Diplomatic Perspectives*. Springer.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Wu, S. Manber, U., Myers, G. and Miller, W. (1990). 'An O(NP) Sequence Comparison Algorithm'. *Inf. Process. Lett.* 35(6): 317-323.

Hearst, M.A. (1994). 'Multi-Paragraph Segmentation of Expository Text'. *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics*. Las Cruces, New Mexico, pp. 6-9.

McCallum, A. K. (1996). *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering, Computer software*. <http://www.cs.cmu.edu/~mccallum/bow> (accessed 10 May 2009).

Salton, G. and Buckley, C. (1988). 'Term-weighting Approaches in Automatic Text Retrieval'. *Information Processing & Management*. 24(5): 513-523.

PREFUSE, *Interactive visualization software*. <http://prefuse.org/> (accessed 10 May 2009).

Naming the unnamed, speaking the unspoken, depicting the undepicted: ***The Australian Women's Register story***

Evans, Joanne

joanne.evans@unimelb.edu.au,
The University of Melbourne, Australia

Morgan, Helen

helen.morgan@unimelb.edu.au
The University of Melbourne, Australia

Henningham, Nikki

n.henningham@unimelb.edu.au
The University of Melbourne, Australia

Ensuring evidence of women's experiences and contributions to our world are kept for the public record and adequately represented in memory institutions has been a key challenge for many inside and outside of the academy over the last half century. This material is vital in order to continue the work of retrieving women's history from 'the shrouds of silence and obscurity' and 'fill in the blank half of a huge canvas'.¹

Over the past decade, the Australian Women's Archives Project (AWAP) has been developing the Australian Women's Register (<http://www.womenaustralia.info/>) as a central part of its strategy to encourage the preservation of women's archival heritage and to make it more accessible to researchers. The Register is a specialist central access point to information about Australian women and their achievements and the multifarious resources in which varying aspects of their lives are documented. It provides a gateway to archival and published material relating to women held in Australian cultural institutions as well as in private hands. A series of small and large grants have contributed to the development of the content of the Register and the technology in which it is captured, managed and made available to as wide an audience as possible via the Web. The National Foundation for Australian Women,² the community organisation behind the AWAP, plays a significant role in securing project funding, along with driving innovation in its coverage and content.

The latest of these grants, an Australian Research Council Linkage Infrastructure Equipment and Facilities Grant (ARC LIEF) awarded in 2008,

allowed the exploration of the Register as part of a federated information architecture to support historical scholarship in digital and networked environments. It involved the investigation of community based methods for populating the Register, as well as enabling the harvesting of its content into emerging national discovery services. With the National Library of Australia (NLA) as a key industry partner, a mechanism for harvesting Encoded Archival Context (EAC)³ records from the Register was established for incorporation into their exciting new *Trove* discovery service,⁴ using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).⁵

The federated information architecture which such harvesting services make possible is aimed at increasing the productivity of all those associated with the creation, management and use of source material for historical research. As well as fostering the development of complicit systems, it is also about allowing a rich multiplicity and variety of voices to contribute their knowledge to resource discovery systems. It involves scholars' direct participation in resource description frameworks allowing their extensive, intimate and fine grained knowledge of sources and their relationships to areas of study to become part of networked information infrastructure. It also aims to provide a mechanism by which the flow of information about resources in and out of cultural institutions is improved, allowing researchers to discover, explore and make connections between materials held in disparate locations efficiently and effectively, and in turn to feed that knowledge back into the network.

As a pioneering e-Research initiative, the story of AWAP and the *Australian Women's Register* (AWR) offers much insight into the establishment, evolution and sustainability of advanced scholarly information infrastructure to facilitate information intensive collaborative research in the humanities.⁶ It is illustrative of how digital and networking technologies change the roles and relationships of scholars, information professionals, universities and the wider community in order to build greater capabilities, connectedness, robustness and resilience into historical/archival/humanities information systems. Above all it asserts the value of scholarly principles, re-visioned, re-imagined and re-distributed for the digital and networked age, and it places women's history firmly in the mainstream rather than being consigned to the margins. What began ten years ago as a small, community initiative aimed at securing the uncertain future of women's archival records has developed into a project of national significance. The fact that it is a feminist project is entirely relevant to the story as well, given the distributed and partial nature of women's

archival collections and the historical circumstances of their production.

This paper will outline and review the development of the *Australian Women's Register*, by discussing the problem of female under-representation in the archival record, explaining the implications of this for historical researchers and describing how the AWR works to harness information about existing records while it creates a new 'community' archive in cyberspace. There will be an emphasis on how it has and has not been able to address emerging requirements for e-Humanities infrastructure as articulated in reports such as *Our Cultural Commonwealth*; however, the focus will be on explaining how the successful development of any e-Humanities infrastructure is shaped by the strength of the collaboration between users and developers.⁷ It will discuss the content and technological developments undertaken as part of the ARC LIEF project, and reflect on the readiness of various stakeholders of the Register to take advantage of these capabilities and participate in the design and development of future ones.

Notes

1. The title of this paper owes much to the wonderful words of Australia's first female Governor General, Quentin Bryce, when re-launching the *Australian Women's Register* on the 13 October 2009. In her speech she highlighted the words of Adrienne Rich, American poet and feminist, 'Whatever is unnamed, undepicted in images, whatever is omitted from biography, censored in collections of letters, whatever is misnamed as something else, made difficult-to-come-by, whatever is buried in the memory by the collapse of meaning under an inadequate or lying language – this will become, not merely unspoken, but unspeakable.' See <http://www.governorgeneral.gov.au/governorgeneral/speech.php?id=625>. The words in quotes come from the same source.
2. Information about the aims of the National Foundation for Australian Women can be found at <http://nfw.org/>.
3. Encoded Archival Context – Corporate bodies, Persons, and Families (EAC-CPF) is a metadata standard for the description of individuals, families and corporate bodies which create, preserve, use, are responsible for, or are otherwise associated with records. Its purpose is to standardize the encoding of descriptions of agents and their relationships to resources and to one another, to enable the sharing, discovery and display of this information. See <http://eac.staatsbibliothek-berlin.de/>
4. *Trove* is the National Library of Australia's new discovery service, providing a single point of access to resources held in Australia's memory institutions and incorporating rich contextual metadata from a variety of sources. See <http://trove.nla.gov.au/>.
5. OAI Protocol for Metadata Harvesting (OAI-PMH) is a lightweight harvesting protocol for sharing metadata between services developed by the Open Archives Initiative. It defines a mechanism for harvesting metadata records from repositories based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language) in support

- of new patterns for scholarly communication. See <http://www.openarchives.org/pmh/>.
6. Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure and the Internet*, MIT Press, Cambridge Massachusetts, 2007.
 7. American Council of Learned Societies, *Our Cultural Commonwealth: The Final Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences*, 13 December 2006, 43 pp, <http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf>.

The Social Lives of Books: Mapping the Ideational Networks of Toni Morrison

Finn, Edward

edfinn@stanford.edu

Department of English, Stanford University, USA

1. Overview

This paper is a case study that is part of a larger Ph.D. dissertation project: an exploration of the networks of references and ideas that make up the social lives of books online. In a time of rapidly evolving ecologies of reading and writing, I argue that the Internet affords us massive amounts of new data on previously invisible cultural transactions. New architectures for reviewing, discussing and sharing books blur the lines separating readers, authors and critics, and these cultural structures capture thousands of conversations, mental connections and personal recommendations that previously went unrecorded. I call these webs of references, allusions and recommendations ideational networks. Using Toni Morrison's career as a model, I will closely examine the ideational networks surrounding her work using the methodologies of social network analysis in order to define a new, statistically informed conception of cultural capital in the digital era.

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 first books, authors, characters and other literary entities (these are the nodes), and second the references linking them together as collocations in book reviews, suggestions from recommendation engines, and other architectures of connection. Advancing from my first case study, Thomas Pynchon, to Morrison's ideational networks, I have discovered the utility of social network analysis methodologies to better analyze graphs of these

literary references. With new data and new tools, I hope to trace the networks of influence and exchange that have contributed to making Toni Morrison arguably the most critically acclaimed and popularly successful author in the United States.

3. Proposal

This paper will present my research on the ideational networks surrounding the works of Toni Morrison. Morrison makes an excellent subject for this kind of study for a number of reasons. As a Nobel laureate and widely read popular author, she has attracted millions of devotees. In her writing she often draws on the African American literary tradition of the talking book, and throughout her career she has explored the ontological power of narrative to create and destroy worlds. It is not surprising, then, that as an author she is deeply committed to expanding the act of reading not only to include millions of people, especially women, who never considered themselves readers before, but also to changing its definition to include conversation, community, and a kind of collaborative reflection. This element of Morrison's authorial appeal is best exemplified by her long-running association with Oprah's Book Club, an enterprise that has had a huge impact on the U.S. publishing industry and on conceptions of reading as a social act.

My presentation will explore the communities of readership that have emerged around Morrison's work and consider the literary company in which her readers and reviewers perceive her. Focusing on a limited set of professional book reviews, reader reviews and recommendations from a dataset of print media, Amazon and LibraryThing, I will map out connections that reviewers and consumers have made between Morrison's works and other literary figures and texts (see Figure 1). I believe these connections will delineate Morrison's position as an extremely popular author who nevertheless challenges her readers to grapple with unflinching, emotionally raw narratives. Her books have introduced millions to deeply troubled corners of American history, combining a modernist style with diverse literary traditions in a way that is both acutely culturally specific and universally compelling. I hypothesize that these factors have driven her remarkable ability to create togetherness and communities of readership even as she traces out the wounds and scars of division, inequality and bias latent in American culture. I also hope to contrast her ideational networks with those of Thomas Pynchon, who has pursued a radically different literary approach through his aversion to publicity and his recondite fiction.



Figure 1: Sample image from work in progress of a Morrison ideational network based on Amazon's recommendation engine.

Here the nodes are books connected together by Amazon's "Customers who bought this also bought" feature, centered on Morrison's novels in the middle. Arrows indicate direction (i.e. Twain's *Huckleberry Finn* is recommended from O'Brien's *The Things They Carried*, but not vice versa). Note both the range of texts and the cultural vectors present, with syllabus classics like Salinger, Steinbeck, Miller and Hawthorne moving down from the center, canonical Native American writers at the top left, etc. Visualization based on Prefuse Java Toolkit.

4. Methodology

This argument will draw on results from several specific datasets of ideational networks.

I have collected professional book reviews from a set of major U.S. newspapers and magazines that consistently reviewed Morrison's publications. These will be analyzed along with customer reviews from Amazon's product pages for Morrison's works, which have been accumulating reviews since 1996. Employing the MorphAdorner project's Named Entity Recognition tool, I am assembling a dictionary of proper nouns that reviewers use as literary references in discussing Morrison's work. Tagging those references in the reviews, I will then explore collocations of references to construct network graphs of the books, authors and other literary entities that reviewers link together.

I have also assembled a database of book recommendations using Amazon's "Customers who bought this also bought" engine and LibraryThing's recommendation engine. These links provide a valuable counterpoint to those works that reviewers choose to mention, since these recommendations are generated by indirect user actions (i.e. when a user buys, reviews or catalogs multiple texts and thereby creates a statistical association among them). Recommendation engines attempt to mimic or track the sale and ownership of cultural products, creating a feedback loop of cultural consumer desire, while review analysis explores a more abstract realm of ideational exchange.

Using methodologies of social network analysis, I will identify those works and authors with the most prestige (i.e. the books most frequently recommended) and centrality (i.e. the author who is best-connected to other authors) in these networks and consider the role of Morrison's texts as centers of ideational networks and, potentially, as bridges between different genre or category groupings. I also hope to explore the role of clustering effects in these networks to see if they are based on predictable factors like genre.

Depending on the speed of my progress with the objectives above and the cooperation of Oprah's Book Club, I also hope to explore the networks of discussion and dialog that have emerged around Morrison's long collaboration with Oprah Winfrey, which has inspired millions of people to take up or return to reading as a leisure activity in adult life. I hope to apply similar methodologies of literary reference to see how Book Club participants contextualized Morrison.

5. Conclusion

As I continue to refine my understanding of ideational networks and improve the methodologies necessary to study them, I am beginning to develop techniques that can effectively be applied to very different authors and provide comparable data. This second case study will provide fertile ground for exploring Toni Morrison's unique authorial fame and to map out the kinds of cultural production that large groups of committed readers can engage in online.

References

- Bourdieu, P.** (1993). *The Field of Cultural Production: Essays on Art and Literature*. Johnson, Randal (ed.). New York: Columbia University Press.
- English, J.** (2005). *The Economy of Prestige*. Cambridge, MA: Harvard University Press.
- Farr, C. K.** (2005). *Reading Oprah: How Oprah's Book Club Changed the Way America Reads*. Albany: State University of New York Press.
- Heer, J.** (2007). *Prefuse Java Toolkit*. Berkeley Institute of Design.
- 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.

Codifica digitale e semiotica della cultura: un esperimento

Fiormonte, Domenico

fiormont@uniroma3.it

Università Roma Tre

Guadalupi, Laura

laura.17@libero.it

Università Roma Tre

1. Inquadramento teorico

L'incontro fra semiotica e teoria della rappresentazione del testo digitale è stato più volte annunciato (Andersen, 1997; Piez, 2007), ma fino ad oggi non si è mai concretizzato in analisi e proposte concrete. È tempo di riprendere in mano la questione e in questo contributo lo faremo partendo da alcuni presupposti teorici per poi illustrare un caso di studio: la traduzione inglese-italiano e poi la codifica XML di una ballata folk irlandese. Come vedremo nell'esempio la pratica delle traduzione interlinguistica e quella della rappresentazione digitale presentano vari punti in comune.

La codifica, al pari della traduzione, è un atto interpretativo (vedi schema sotto). Se tradurre interlinguisticamente significa dare forma alla realtà attraverso una lingua naturale che è, nelle parole di Lotman, un 'sistema modellizzante primario' (Lotman, 1970), codificare è un'attività analoga, che coinvolge un altro sistema di segni. Nel caso da noi preso in esame allora il linguaggio di markup può essere considerato un 'sistema modellizzante secondario' (cf. Fiormonte, 2008: 294-295). L'XML, metalinguaggio descrittivo e dichiarativo, si affrancia quindi dalla tradizionale concezione semiotica che relega il medium di comunicazione a un ruolo marginale, per diventare anch'esso strumento produttore di senso. Il documento digitale infatti, e ancora di più un insieme strutturato di documenti digitali – biblioteca, mediateca, ecc. –, si inscrive nel dominio delle auto-descrizioni di una cultura: 'La differenza essenziale tra l'evoluzione culturale e l'evoluzione naturale sta nel ruolo attivo delle autodescrizioni, nell'influenza esercitata sull'oggetto dalle rappresentazioni dello stesso' (Lotman and Uspenskij, 2006: 152). Ciò vuol dire, nel nostro caso, che il documento digitale svolge un doppio ruolo di testimone attivo: sia come 'documento' (che può incarnarsi in diversi formati (PDF, HTML, ecc.) sia in quanto 'rappresentazione' (cf. Buzzetti, 2006: 55). Ci sembra

chiaro quali siano le influenze di una siffatta procedura, giacché la codifica digitale è un tipo di auto-descrizione che è in grado di 'definire' la natura di un documento nel momento in cui lo 'descrive'.

Riassumendo dunque la codifica digitale di un documento può essere considerata un *knowledge-shaping process* costituito da quattro dimensioni interagenti:

- **Trascrizione;** processo di selezione 'relying on conventions and reflecting theoretical goals and definitions' (Duranti, 2006: 302);
- **Traduzione;** processo di appropriazione e adattamento al sistema semiotico di arrivo (Lotman, 1985: 113-129);
- **Interpretazione;** processo ermeneutico di creazione di nuova conoscenza attraverso strumenti di esplicitazione formale (Orlandi, 1990: 26-27);
- **Modellizzazione;** qui il termine può essere inteso in una duplice accezione: a) in senso semiotico-culturale come processo di costruzione di *device* meta culturali (Lotman and Uspenskij 2006); b) come processo di costruzione di un modello euristico del documento, operazione preliminare a qualsiasi trattamento informatico (Gigliozi, 1997).

Nessuno di questi livelli può essere considerato a sé e l'ordine di questa lista risponde solo a una comodità espositiva.

2. Traduzione e codifica come processi culturali

Lingue e testi interagiscono all'interno di quel delicato processo generatore di senso che è la traduzione e che nella prospettiva indicata da Lotman consideriamo come concetto di 'confine', luogo di trasmissione e di trasformazione della memoria. Come scrive Lotman, l'attività culturale consiste nel 'tradurre un certo settore della realtà in una delle lingue della cultura, trasformarlo in un testo, cioè in una informazione codificata in un certo modo, introdurre questa informazione nella memoria collettiva' (Lotman, 1970; trad. it. 1975, 25-35).

Partendo da tali premesse, abbiamo sviluppato l'analisi in due prospettive convergenti.

La prima ci ha portato a elaborare una traduzione interlinguistica dall'inglese all'italiano di una ballata folk irlandese sull'Easter Rising (cfr. frammento in Tabella 1), la rivolta del 1916 contro il dominio britannico. Al livello macro-strutturale dei rapporti fra i due sistemi-lingua, abbiamo optato per una traduzione *source oriented*, che conservasse

le differenze culturali, senza quindi incorrere in distorsioni. Sottoscriviamo l'idea di Jurj Lotman e Peter Torop (Lotman, 1984; trad. it. 1985; Torop, 1995) sulla traduzione come appropriazione di una cultura altra, come processo di arricchimento, di integrazione, piuttosto che di assorbimento livellante. Al livello micro-strutturale dell'analisi semiotica del prototesto abbiamo applicato le nozioni della semiotica culturale, in particolare i concetti di 'semiosfera' e 'confine', mettendo in luce le dinamiche conflittuali che possono scaturire quando due culture entrano in concorrenza per affermare ciascuna il proprio sistema di senso. La semiosfera irlandese (vedi Fig. 1; tag 'sem-irl') e quella britannica (Fig. 1; tag 'sem-brit') sono cerchi dalla cui intersezione scaturisce la sfera semiotica della guerra Anglo-Irlandese, delimitata dalle figure di confine che sono i protagonisti della *ballad*, ossia i rivoluzionari che combatterono per l'indipendenza.

Ritornello tratto dalla versione originale della ballad <i>Freedom's Sons</i>	Versione tradotta <i>I Figli della Libertà</i>
<i>They were the men with a vision, the men with a cause The men who defied their oppress's laws The men who traded their chains for guns Born into slav'ry, they were Freedom's Sons</i>	Erano uomini con un sogno, uomini con un motivo per combattere Uomini che sfidarono le leggi dei loro oppressori Uomini che barattarono le loro catene con le pistole Nati schiavi, erano i Figli della Libertà

Tabella 1. Frammento della ballata Easter Rising con relativa traduzione.

Il secondo passaggio è consistito nella traduzione metalinguistica, ovvero la codifica della ballata usando lo schema XML-TEI (Burnard and Sperberg-McQueen 1995). A monte del lavoro del filologo digitale stanno le scelte pragmatiche circa gli elementi testuali da marcare, e il linguaggio di markup obbliga a esplicitare tali scelte, motivandole. In virtù di tali analogie, per indicare il processo di digitalizzazione, possiamo adoperare l'etichetta di *traduzione di secondo livello*. Al posto della codifica delle fasi di stesura, con tanto di analisi critica alle modifiche operate dall'autore della canzone, si è optato per una codifica della canzone che marcasce le nostre interpretazioni del testo. Abbiamo così adattato la DTD della TEI inserendo all'interno dei tag che descrivono la struttura sintattica un valore che rimanda alla funzione narrativa dell'occorrenza e alla semiosfera cui essa appartiene. Per esempio nella Fig. 1., rigo 15, 'ana= fig-enf sem-irl sem-rising ref-prop', riferito a *the men*, dichiara la natura retorico-enfatica dell'occorrenza e la sua funzione narrativa di protagonista della diegesi. Inoltre, il valore 'sem-irl' rimanda alla semiosfera di appartenenza irlandese, mentre 'sem-rising' sta ad indicare che i rivoluzionari sono i promotori della lotta (la terza sfera di senso) e al contempo agiscono da traduttori sul confine poroso delle due macro-sfere intersecanti.

Per converso, al rigo 17, troviamo il riferimento agli *oppressors*, antieroi e protagonisti della semiosfera britannica. In conclusione dunque si modifica il ruolo del codificatore: da traduttore delle scelte autoriali nell'ambito del processo di scrittura, a co-artece del testo, poiché nella codifica immette la sua personale interpretazione del brano.

```

<interp&gt; type="referente" resp="editor">
  <interp id="ref-prost" value="protagonista-eroe"/>
  <interp id="ref-ant" value="antagonista-antieroe"/>
  <interp id="Pearse" value="poet"/>
</interp&gt;
<interp&gt; type="semiosfera" resp="editor">
  <interp id="sem-irl" value="libertà"/>
  <interp id="sem-brit" value="schiavitù"/>
  <interp id="sem-rising" value="lotta"/>
</interp&gt;
<lg id="L02" ceh="italic" type="fig-presa-seconda-quartina">
  <l id="12_1"><seg id="12_2" ana="ref-poet sem-irl sem-rising">They</seg>&gt;<seg id="12_3" ana="fig-enf sem-irl sem-rising ref-prop">the men</seg>&gt;with a vision,<seg id="12_4" ana="fig-enf sem-irl ref-prop">the men</seg>with a cause</l>
  <l id="12_5" ana="fig-enf sem-irl sem-rising ref-prop">The men who defied their <seg id="12_6" ana="sem-brit ref-ant">oppressors</seg>&lt;seg id="12_7" ana="sem-brit">laws</seg><l id="12_8" ana="fig-enf sem-irl sem-rising ref-prop">The men who traded their <seg id="12_9" ana="fig-sim sem-brit">chains</seg>&lt;seg id="12_10" ana="sem-rising">guns</seg><l id="12_11" ana="fig-sim sem-brit">Born into <seg id="12_12" ana="fig-sim sem-brit">slav'ry</seg>,<seg id="12_13" ana="sem-irl sem-rising ref-prop">they</seg>&gt;were<seg id="12_14" ana="fig-ant fig-sim sem-irl">Freedom's Sons</seg></l>

```

Figura 1

È questa solo una prima ipotesi, utile nel caso in cui di un oggetto di studio si intendano offrire tante codifiche quante sono le prospettive di analisi adoperate, le figure di specialisti coinvolte, ecc. A una digitalizzazione del processo compositivo tradizionalmente diacronica, verrebbe quindi accostata una resa metalinguistica delle sue molteplici interpretazioni sull'asse sincronico. Ciò per sottolineare, ancora una volta, la natura poliedrica, extrasemiotica ed intersemiotica di qualsiasi proto- e meta-testo (cfr. Bachtin, 1979; trad. it. 1988, 292-293).

3. Conclusioni. La pluridimensionalità semiotica della codifica

La trascrizione e codifica di un documento in XML produce un proprio e indipendente 'strato' semiotico che si affianca e sovrappone alla fonte. Realizzare tali strati attraverso un linguaggio formale implica delle scelte da parte del codificatore-traduttore-interprete che, non differentemente da un parlante di una lingua naturale, si trova continuamente di fronte a diverse possibilità di 'interpretare' e 'rappresentare' lo stesso oggetto. Ma come avviene questo slittamento e insieme 'incremento' semiotico? XML, in particolare, costituisce un ulteriore passo avanti verso il concetto di metalinguaggio (un linguaggio che 'parla' di un altro linguaggio), ma anche un potente strumento in grado di generare, a partire da una medesima sintassi, altri linguaggi fratelli, nonché diverse ipostatizzazioni espressive dello stesso contenuto.

Il computer è un 'symbolic tool' (Andersen, 1997: 1) ed è dunque tempo di iniziare ad analizzare in termini semiotici tanto i prodotti che i processi

della digitalizzazione. I linguaggi di codifica, nel caso analizzato lo schema XML-TEI, possono essere considerati vere e proprie 'metalingue' capaci di rappresentare e tradurre la conoscenza. In accordo con la visione della *cultural informatics* proposta da Crane et al. (2008) la semiotica può esserci di aiuto per definire le basi di una teoria culturale della codifica digitale, ovvero dei modi e delle procedure in cui gli strumenti di digitalizzazione sviluppano, traducono e modificano i meccanismi della memoria e delle identità culturali.

References

- Andersen, P.B.** (1997). *A Theory of Computer Semiotics. Semiotic Approaches to Construction and Assessment of Computer Systems*. Cambridge: Cambridge University Press.
- Bachtin, M.** (1979). *Estetika slovesnogo tvorcestva*, Izdatel'stvo «Iskusstvo». Torino: Einaudi [trad. it. 1988, L'autore e l'eroe. Teoria letteraria e scienze umane].
- Buzzetti, D.** (2006). 'Biblioteche digitali e oggetti digitali complessi. Esaustività e funzionalità nella conservazione'. *Archivi informatici per il patrimonio culturale*. Roma: Bardi Editore, pp. 41-75.
- Crane, G., Bamman, D. and Jones, A.** (2008). 'ePhilology: When the Books Talk to Their Readers'. *A Companion to Digital Literary Studies*. Siemens, R., Schreibman, S. (ed.). Oxford: Blackwell. <http://www.digitalhumanities.org/companionDLS>.
- Duranti, A.** (2006). 'Transcripts, Like Shadows on a wall'. *Mind, Culture, Activity*. 13(4): 301-310.
- Fiormonte, D.** (2008). 'Il testo digitale: traduzione, codifica, modelli culturali'. *Italianisti in Spagna, ispanisti in Italia: la traduzione. Atti del Convegno Internazionale*. Piras, P. R., Alessandro, A., Fiormonte, D. (ed.). Roma: Edizioni Q, pp. 271-284.
- Gigliozi, G.** (1997). *Il testo e il computer. Manuale di informatica per gli studi letterari*. Milano: Bruno Mondadori.
- Lotman, J. M.** (1984). 'O semiosfere'. *Töid märgisüsteemide alalt / Sign Systems Studies / Trudy po znakovym sistemam*. 17: 5-23, [trad. it. 1985, *La semisfera. L'asimmetria e il dialogo nelle strutture pensanti*, a cura di S. Salvastreni, Venezia, Marsilio].
- Lotman, J. M.** (1985). 'Una teoria del rapporto reciproco fra le culture'. *La semiosfera. L'asimmetria e il dialogo nelle strutture pensanti*. 113-129.
- Lotman, J. M. and Uspenskij, B. A.** (1970). 'Kul'tura i informacija'. *Stat'i po tipologii kul'tury. Materialy k kursu teorii literatury*. Tartu [trad. it. 1975, Tipologia della cultura, Milano, Bompiani, pp. 25-35].
- Lotman, J.M. and Uspenskij, B. A.** (2006). 'Eterogeneità e omogeneità delle culture. Postscriptum alle tesi collettive'. *Tesi per una semiotica della cultura*. Lotman, J.M. (ed.). Roma: Meltemi, pp. 149-153.
- Orlandi, T.** (1990). *Informatica umanistica*. Roma: La Nuova Italia Scientifica.
- Piez, W.** (2007). 'Form and Format: Towards a Semiotics of Digital Text Encoding'. *Digital Humanities 2007 Conference Abstracts*. http://www.digitalhumanities.org/dh2007/abstracts/paper_188_piez.pdf.
- Sperberg-McQueen, C. M. and Burnard, L.** (1995). 'TEI Lite: An Introduction to Text Encoding for Interchange'. *TEI U5*. Text Encoding Initiative. <http://www.tei-c.org/TEI/Lite>.
- Torop, P.** (1995). *Total'nyj perevod*. Tartu: Tartu Ulikooli Kirjastus [trad. it., 2000, *La traduzione totale*, a cura di B. Osimo, Rimini, Guaraldi].

Open vs. Closed: Changing the Culture of Peer Review

Fitzpatrick, Kathleen

kfitzpatrick@pomona.edu

Pomona College, USA

Despite our differing research methodologies, subjects, and motives, the one thing scholars across the disciplines and around the world might agree upon is the significance of peer review. Peer review may be the sine qua non of academic work; we use it in almost everything we do, including grant and fellowship applications, hiring and promotion processes, and, of course, in vetting scholarly work for publication. We all operate under the agreement that peer review is a good thing, by and large, both a means of helping scholars improve their work and a system for filtering that work for the benefit of other scholars.

However, as I argue in *Planned Obsolescence*, the means by which we conduct peer review demand careful reconsideration and revision as academic publishing moves increasingly online. Clay Shirky has argued that the structures of the internet demand a “publish, then filter” process, encouraging the open communication of *all* of the results of scholarly investigation, followed by a process that filters those results for quality (Shirky 2008). I explore the reasons such a transformation is desirable at length in *Planned Obsolescence*, primarily that it makes little sense to replicate a mode of review designed for print’s economics of scarcity within the internet’s economics of abundance (see Jensen 2007); if what is scarce in the age of the network is not the means of production but the time and attention available for consumption, the best use of peer review would be to help researchers find the right text, of the right authority, at the right time.

A born-digital system of review would work with rather than against the strengths and values of the network by privileging the open over the closed, and by understanding the results of peer review as a form of metadata enabling scholars to find and engage with research in their fields. How to build and implement such a system, however, remains in question: how do we devise a networked review system that is open, honest, and thorough, that draws the best from the “wisdom of the crowds” (Surowiecki 2004; Anderson 2006) while upholding the standards that review is meant to serve?

Several examples of online review processes already exist; within humanities journal publishing, the most

significant may be that of *electronic book review*; articles submitted there are posted in a password-protected review space, where registered users of the site can read them, leave glosses, and recommend acceptance or rejection. However, though the editors use my term “peer-to-peer review” in describing their system, it falls a bit short of the truly open system I imagine; the review system is still kept behind the scenes, and while the reviews are crowd-sourced, the reviewers producing them aren’t asked to take responsibility for their opinions by expressing and defending them in public. This aspect of peer-to-peer review is key; just as the quality of the algorithm determines the quality of a computational filtering system, the quality of the reviewers will determine the quality of a human filtering system. Online peer review must made open and public not just as a means of increasing communication but as a means of increasing reviewer accountability, providing for the ongoing review not just of texts but of reviewers.

In order to experiment with the possibilities for an open review system, and with the consent of NYU Press and my editor, Eric Zinner, I placed the entirety of *Planned Obsolescence* online in late September 2009. The text was published in CommentPress, a WordPress plugin developed by the Institute for the Future of the Book, which enables the discussion of texts at a range of levels of granularity, from the paragraph to the page to the document as a whole. At the same time, NYU Press sent the project out for traditional peer review.

Such experiments have been conducted before; in 2008, Noah Wardrip-Fruin published a draft of *Expressive Processing* through *Grand Text Auto*, while MIT Press sent it to outside readers. Noah, however, wasn’t seeking to create a head-to-head contest between closed and open review; he was motivated by the desire for feedback from a community he trusted (see Wardrip-Fruin 2009b). My motives were a bit more complex; I wanted that same community-based feedback, but I also wanted to test open review against more traditional reviews, to gauge differences in the kind and quality of responses produced within an online system, and to project the kinds of changes to CommentPress that might help transform the plugin into a viable mechanism for peer-to-peer review.

In slightly less than six months, *Planned Obsolescence* received 205 comments from 39 different readers (not counting my own 78 responses). These comments are by and large excellent, and have been extremely helpful in thinking about the revision of the manuscript. Most of the comments, however, are locally oriented; CommentPress’s paragraph-level commenting strategy encourages a close attention to the particulars of a text rather than a more holistic

discussion. This focus on the text's details in the comments wasn't unexpected; we anticipated that the traditional reviews, being written after the entire manuscript had been read, would tend to focus a bit more on the big picture than would comments made in *medias res*. This assumption did largely bear out; the offline reviewers tended more toward an assessment of the overall argument.

Our first tentative conclusion, then, was that a functional open review system would require clearer ways for online reviewers to leave broader comments. An update to the CommentPress plugin, released a couple of months into our experiment, helped provide that functionality by highlighting the "community blog" section of the site, which in theory would allow members of a community of readers to engage one another in discussion of their reviews and of the project as a whole. In actual practice, however, that engagement did not occur, though it remains unclear whether this is due to the blog's belated introduction or some other issue.

Additionally, however, Zinner asked the offline reviewers whether they would be willing to participate in our process, allowing us to post their reviews for discussion and response; one, Lisa Spiro, agreed. Spiro's willingness to participate, and the generosity of her review, revealed the importance of the social commitments involved in the peer review process. Those scholars who have long undertaken the often thankless work of peer review have largely done so out of a commitment to the advancement of knowledge in the field. But fostering participation in online discussion requires not just intellectual interest on the part of individuals but also a solid, committed social network. Reviewers participating in an open process must have a stake in that process beyond that of the disinterested reader; they must understand the text and its author to be part of a community in which they are invested and to which they are accountable.

Beginning in March 2010, MediaCommons will conduct another open review experiment, publishing a small group of papers being considered for a special issue of *Shakespeare Quarterly*. Through this experiment, we hope to explore a number of variables: the relative weights of commitment to subject matter and commitment to digital methodologies in determining participation in open review, the level of engagement in the review of article-length (as opposed to book-length) texts online, and the structures of participation in the review of work by multiple authors in one venue.

Both experiments involve the review of comparatively traditional forms of scholarship, the book and the journal article, which we have opted to begin with for two reasons: first, that transforming

the processes of reviewing these forms of scholarship presents the broadest potential impact on academic publishing as it exists today, and second, that it confines the question under consideration to *mode* of review, rather than expanding into *criteria* for review. That last is extremely important; many, if not most, scholars working in new forms of multimodal scholarship have encountered the sense that the academy in general does not know how to review such work. We hope to experiment in the future with models for review of new forms of scholarship.

This paper will, in the end, argue that a truly effective peer-to-peer review system will need to place its emphasis not just on developing the technological network but on developing the social network; it must be focused around clusters of scholars who are already in dialogue with one another. It must also be accompanied by a shift in values that encourages scholars to understand the business of reviewing as being a commitment not just to the advancement of intellectual thought but to the structure of that community and its dialogue. And, as participation in such review requires significant time and energy from a larger number of scholars than traditional review, the academy must recognize the importance of reviewing, acknowledging the significant labor involved, creating structures through which reviewers can receive "credit" for their work.

MediaCommons hopes to foster these developments in a genuinely peer-to-peer review process in the coming months by adding functionality allowing readers to rate and respond to the comments left by others, as well as by building links between these reviews and our social network (see Fitzpatrick 2009b), making a scholar's reviews visible as a part of their portfolio of scholarly work. Together, we hope, these links will allow peer review to become an open, social process, one that will transform online peer review into a mechanism for collaborative post-publication filtering, helping authors to improve their work and enabling researchers to find and assess the authority of the texts they need, by working with rather than against network-based interactions.

References

- Anderson, C.** (2006). 'Wisdom of the Crowds'. *Nature*. <http://www.nature.com/nature/peerreview/debate/nature04992.html>.
- Fitzpatrick, K.** (2009a). 'Planned Obsolescence: Publishing, Technology, and the Future of the Academy'. *MediaCommons*. <http://mediacommons.futureofthebook.org/mcpress/plannedobsolescence/>.

Fitzpatrick, K. (2009b). 'MediaCommons as Digital Scholarly Network: Unveiling the Profile System'. *MediaCommons*. <http://mediacommons.futureofthebook.org/blog/2009/11/10/mediacommons-digital-scholarly-network-unveiling-profile-system>.

Jensen, M. (2007). 'The New Metrics of Scholarly Authority'. *Chronicle of Higher Education*. <http://chronicle.com/article/The-New-Metrics-of-Scholarly/5449>.

Shirky, C. (2008). *Here Comes Everybody: The Power of Organizing Without Organizations*. New York: The Penguin Press.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. New York: Doubleday.

Tabbi, J. (ed.) (n.d.). *Electronic book review*. <http://electronicbookreview.com/>.

Wardrip-Fruin, N. (2009a). *Expressive Processing: Digital Fictions, Computer Games, and Software Studies*. Cambridge, Mass: MIT Press.

Wardrip-Fruin, N. (2009b). 'Blog-Based Peer Review: Four Surprises'. *Grand Text Auto*. <http://grandtextauto.org/2009/05/12/blog-based-peer-review-four-surprises/#0>.

Using ODD for Multi-purpose TEI Documentation

Flanders, Julia

Julia_Flanders@brown.edu

Women Writers Project, Brown University, USA

Bauman, Syd

Syd_Bauman@brown.edu

Women Writers Project, Brown University, USA

The philosophy of "literate programming" (Knuth 1984), on which the TEI ODD is founded, proposes that code and documentation be written and maintained as a single integrated resource, from which both working programs and readable documentation can be generated. As currently designed, the TEI ODD system supports these goals, and there exist several good examples of extended project documentation written using the ODD customization file (see (Trolard 2009), (Burnard and Sperberg-McQueen 2006), (Burnard et al. 2010) and also the TEI Guidelines themselves). However, at present these examples only assume and demonstrate the ability to generate two types of documentation: a prose narrative and a set of reference documentation. As text encoding projects develop and mature they generate a variety of documentation that may include training tutorials and reference documentation, public documentation of editorial and encoding practices, documentation of their TEI customization, and internal documentation of the encoding decisions that have resulted in their current encoding rationale. Many of these other forms of output (such as training tutorials) have not been tried in practice and the current ODD processor, Roma, does not explicitly support them. In this paper we explore the possibility of generating more complex and varied forms of documentation using the TEI ODD customization file.

As background for this discussion we should begin by describing the nature and role of this customization file. Underlying any TEI-encoded document is a schema defining the terms of its validity, and underlying that schema is a further specification: the ODD customization file, which is the source file from which the schema is generated (documented in detail in chapters 22 and 23 of the TEI Guidelines (TEI 2007)). The ODD file serves several important functions:

1. To express the specific choices that are being made with respect to the TEI system as a whole: which TEI modules are to be included in the generated

schema, which elements and attributes from these modules are to be included or omitted, changes to content models, etc.

2. To document those choices: for instance, to explain the meaning of controlled vocabularies for attribute values, or to express the rationale for applying an element only in specific contexts or with a slightly broader or narrower significance than described in the Guidelines.
3. To permit the generation (using these two kinds of information) not only of a custom TEI schema but also of custom documentation.

This custom documentation includes a re-expression of the TEI reference documentation: that is, the second volume of the printed TEI Guidelines, the portion containing separate entries for each element, attribute, class, and macro. This custom reference documentation includes only references to elements, attributes, and classes that are actually present in the custom schema, and includes as part of these entries some of the additional documentation expressed as part of point 2 above (e.g. glosses of specific attribute values, etc.).

The goal of the TEI customization mechanism as a whole, then, is two-fold. First, it aims to make TEI schemas self-documenting, by encapsulating all of the choices made in a separate document (the customization ODD file) that can be stored, maintained, exchanged. And second, to a certain extent the mechanism is intended to permit *encoding systems* to be self-documenting, in the sense that the documentation of the encoding practice can be bundled together with the raw materials for creating and maintaining the custom schema. This is true in a very straightforward manner to the extent that information about encoding practice can be directly associated with specific schema modifications. But more complex documentation is also possible: since the ODD file is a TEI document, one can also include more detailed documentation that is not directly associated with a schema modification, simply by adding prose to that TEI document. When the ODD file is processed, the resulting documentation will include that additional prose. This more detailed approach is currently uncommon, but this is primarily because the Roma web interface does not currently support the authoring of such documentation; users need to author the ODD directly in order to create more complex documentation forms. Examples of this more detailed approach include the TEI in Libraries best practices documentation (Hawkins and Bauman 2009), and also the documentation for TEI Lite (Burnard and Sperberg-McQueen 2006).

With this in mind, however, it is tempting to extend this process even a step further: to use the ODD

file as a way of writing documentation of other sorts that are even less closely attached (in their methods of organization) to the schema specifications. For example, for many purposes a project may need to maintain both reference-style documentation for each element or encoding concept, and also tutorial-style documentation whose emphasis is on leading the reader through a pedagogically structured narrative. Encoders learning to transcribe manuscript diaries might need to learn first the specific set of structural elements that will be used to encode the overall structure of the document (<div>, <opener>, <dateLine>, etc.) and then the set of elements having to do with transcriptional difficulties (<gap>, <unclear>, <add>,). At the same time, in other areas of the project it might be essential to have documentation of the underlying rationale for the encoding approach, or a high-level narrative with links to specific entries. More importantly, different tutorials or forms of documentation might need to use particular portions of the specification in different orders: the tutorial format might take specific sets of elements and treat them as groups, while a more comprehensive narrative documentation might treat the same encoding concepts in alphabetical order, or by conceptual grouping, or by TEI module (to take just a few examples).

To support the generation of multiple documentation narratives from a single ODD requires two changes to the way the ODD is written. First, additional prose (from which the various narratives will be generated) will need to be included in the ODD, and the ODD itself may need to be organized somewhat differently to accommodate this prose. Second, and more challengingly, some mechanism is required by which the different narrative orderings can be expressed in the ODD and then processed to produce the various appropriate forms of output. These different forms of documentation could of course be maintained separately (as is currently the case) but the value of the ODD system lies in its philosophy (expressed somewhat obscurely in the word "ODD", or "one document does-it-all") of having a single document that expresses and documents all aspects of a TEI encoding scheme. Rather than creating separate prose documentation for these additional forms, there is clear value in being able to generate multiple forms of documentation serving different purposes, from a single ODD.

We are not aware of any examples of this latter type, but the utility of this approach is clear and the ODD language – because it is part of the TEI and thus can use the full expressive resources of the TEI language as a whole – contains the features necessary to support it. In this paper we present an initial implementation in which we

construct a set of tutorials on specific encoding topics in areas where the WWP has customized the TEI (for instance verse, title pages, and notes), using the ODD mechanism. The approach we explore entails encoding the components of the various narratives using standard TEI prose and documentation elements, and constituting each individual narrative using stand-off markup to identify and assemble the pieces in the appropriate order for the documentary form in question. We present the proposed ODD encoding for these tutorials, using the Women Writers Project internal documentation as a testbed. We also present prototype stylesheets that will generate multiple documentation narratives from the single ODD source.

References

- Burnard et al.**. *An Encoding Model for Genetic Editions*. <http://tei.svn.sourceforge.net/viewvc/tei/trunk/genetic/geneticTEI.xml?view=markup>.
- Burnard, L. and Rahtz, S.** (2002). 'The Role of the TEI in the Authoring and Interchange of XML Documents'. *Proceedings of elpub2002*. Berlin, 2002. <http://elpub.scix.net/data/works/att/02-22.content.pdf>.
- Burnard, L. and Sperberg-McQueen, M.** (2006). *TEI Lite: Encoding for Interchange: an introduction to the TEI*. <http://www.tei-c.org/release/xml/tei/custom/odd/teilite.odd>.
- Hawkins, K. and Bauman, S.. TEI in Libraries**. *Digital Library Federation*, 2009. <http://github.com/sydb/TEI-in-Libraries/tree/master/BestPractices>.
- TEI (2007)** (2007). *Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html>.
- Trolard, Perry** (2009). *TEI Tite: A recommendation for off-site text encoding*. http://www.tei-c.org/release/xml/tei/custom/odd/tei_tite.odd.
- Knuth, Donald E.** (1984). 'Literate Programming'. *The Computer Journal*. **27(2)**: 97-111. <http://www.literateprogramming.com/knuthweb.pdf> <http://doi:10.1093/comjnl/27.2.97>.

Xiakou: A Case Study in Digital Ethnography

Flower, John

FlowerJ@sidwell.edu
Sidwell School, Washington, D.C

Leonard, Pamela

pamleonard@hughes.net
Independent Scholar

Martin, Worthy

wnm@cs.virginia.edu
Institute for Advanced Technology in the Humanities, University of Virginia

This story begins in a small Sichuan village over fifteen years ago as a historian (John Flower) and an anthropologist (Pamela Leonard) began their study of the cultural landscape of a contemporary Chinese village. The story evolves as they strive to pioneer *digital ethnography* and later, in collaboration with The Institute for Advanced Technology in the Humanities (IATH), build interactive presentation of focused, long-term fieldwork research results in the form of an online monograph, media archive, and information repository, entitled *Moral Landscape in a Sichuan Mountain Village: A Digital Ethnography of Place*.

The original and ongoing field study explores the histories, beliefs, livelihoods, and local identities in Xiakou Village, located in the mountains of Ya'an County, in western Sichuan Province of the People's Republic of China. The goal of the project is to understand Xiakou Village as an evolving *cultural landscape*, defined as the interwoven field of physical environment, historical memory, and moral agency, in which particular places gather a people's sense of themselves and serve as sources of belonging and identity. This understanding attempts to establish a basis to consider questions such as: What does it mean to belong in a place? How do people understand who they are in terms of where they live? What is the relationship between history and place? How do memory and landscape inform the ways in which people define their communities?

The ethnography uses the new possibilities of digital technology to create interleaving essays, primary source multimedia artifacts, and GIS maps. The purpose of this digital form is to render more transparent the relationship between source and interpretation, to open up non-linear narrative paths through the ethnography, and therefore to more vividly reveal the interconnections among different dimensions of village life that are the core content of

the project. Indeed, we revisit the village study model to highlight the overlapping fields of interaction that link the village to broader regional, national, and even transnational identities.

Another fundamental aim of the project is to reframe modern Chinese history away from the big narrative of the nation and toward local stories of the grassroots. How do the villagers of Xiakou understand their history? What memories and meanings from the past still animate their place, and how are they remembered and explained?

Moral Landscape in a Sichuan Mountain Village is multidisciplinary, using the perspectives of history, anthropology, economics, folklore, and religion to try to understand the interconnected facets of life expressed in the village landscape. The common thread running through the ethnography is the idea that the landscape holds moral values. When people in Xiakou talked about place and history they were talking about what was good and bad, right and wrong.

We understand digital ethnography to be an online interactive monograph with integrated archive and database. This digital format of the ethnography evokes an understanding of place through interactive essays that localize the broad trends of China's modern history in the lived experiences of Xiakou's villagers. The interactive essays are the project's main narrative tissue, interconnected by a searchable archive of digital artifacts. These artifacts consist of multimedia information—photographs, scanned documents, audio and video recordings, GIS maps—contextualized in a thick setting of related metadata, and shared across essays. The project's digital format is essential for realizing the rich potential of the ethnographic and historical content of the research: a central database and interconnected XML content enable the transparency, connectivity, and interactivity that comprise the key innovative characteristics of this form of narrative. Transparency means that the ethnography will reveal not simply "what we know" but also "how we know it," by providing the reader access to primary source materials in the database. The architecture of the interactive interface will also use the database to encourage connections across thematic categories, making it possible for the reader to explore alternatives to a set, linear narrative.

2. Ways of belonging: new village studies and mapping the cultural landscape

What is a village in China? A wide range of scholarship has addressed this central question, from the perspectives of regional systems analysis

(Skinner 1964) to cultural landscape studies (Knapp 1992, Feuchtwang 1997). Our approach tries to give priority to villagers' conscious representations, analyses, and understandings of their relationship to "their place". The resulting geographical scope goes beyond the village itself to encompass the communities along the North Road and, under some conditions, extends to include the broader eight county Ya'an region.

In *Moral Landscape in a Sichuan Mountain Village* we advocate a return to the ethnographic tradition of village studies, but using new tools of the digital humanities that emphasize the ways in which place is not simply a fixed and unchanging location, but rather a nexus of evolving relationships and historical connections to other places. Thus, one of our goals is to highlight the multiple, overlapping fields of interaction that link the village to broader regional, national, and even transnational identities.

We see our project as complementary to the much larger and comprehensive initiatives that aim to create complete datasets, such as the China historical GIS project (Bol 2006). In contrast, our project does not attempt to be comprehensive, but rather celebrates the particularity of place. We hope that our qualitative interpretation of landscape will provide the kind of unique local portrait of place from which comprehensive projects can create a more vivid broad tableau of China as whole.

3. Beyond Revolution: an inductive approach to local history

Another fundamental goal of our project's landscape approach is to reframe modern Chinese history away from the master narrative of the nation and toward local stories of the grassroots (Duara, Prazniak). How do the villagers of Xiakou understand their history? What memories and meanings from the past still animate their place, and how are they remembered and explained? How does that local understanding of history reiterate or differ from historical narratives based on the nation-state, China, as subject? While there are excellent village-based histories (e.g. Chan, Madsen, Unger 1992; Selden, Friedman, Pickowicz 1991) that focus on the local impact of national events, particularly the Chinese revolution, in *Moral Landscape in a Sichuan Mountain Village* we try to adopt a more localized, inductive approach. The historical scope of our project thus largely corresponds to the way villagers mark the turning points in their past, based on their personal experiences in local places and marking events that fall within their horizon of memory.

Methodologically, we understand that the essays and artifacts represent *our synthesis* of a dialogue with local villagers and with local historical source-materials on the topic of social and environmental change. In confronting the subjective reality of fieldwork and analysis, anthropologists have emphasized the need to be transparent in presenting the politics of the research encounter. We believe digital technology allows us to go further in meeting this aim.

4. History, environment, and agency in the moral landscape

In trying to understand the significance of the environmental changes that have taken place in this valley, we frame issues of environment and economic development within local cultural practices and historical knowledge. How do local people draw on their historical understanding of place in adapting to economic development policies introduced from outside? How do those development policies in turn influence their livelihoods, and change their understanding of the landscape?

5. Structure, content, and logic of the digital ethnography

The structure of the ethnography's online monograph comprises eight chapters: History, Landscape, Belief, Folklife, Authority, Work, Gazetteer, and Biography. Chapters are not airtight divisions, but rather groupings that highlight the dominant themes of the essays within them. There are three main types of content within this chapter structure: *essays*, *interactive maps*, and *artifacts*. Essays are the basic interpretive building blocks of the ethnography and are accessed through the chapters. The interactive maps under the Gazetteer chapter will offer spatial representations of sites in the cultural landscape, dynamically presented through GIS layers, sorted by kind and historical period.

Both the maps and essays are illustrated and documented by "artifacts", i.e., foci of evidence that link to multimedia content—photographs, video and audio recordings, image maps, diagrams, supplemental texts, primary source documents, and field notes. The artifact frames this multimedia content within supplementary metadata and highlights thematic overlaps and interconnections within the ethnography.

The essay/artifact structure allows us to experiment with different approaches to conceptualizing and presenting the ethnographic research. These artifact-centered essays are intentional inversions of the more familiar text-driven narrative presentation, and

they point the way to readers who want to engage the ethnography more interactively. To enable that level of engagement, our goal is to code each artifact and each essay subsection with selections from a finite set of keywords, making the whole site fully searchable through the site's integrated information structures.

6. Proposed Presentation

We will discuss the information structures in which the base materials are created and maintained. Then we discuss the interactive interface through which those materials are accessed by scholars and the general public. Finally, we will justify our claim that these techniques embody the methodologies expressed above.

Challenges of Linking Digital Heritage Scientific Data with Scholarly Research: From Navigation to Politics

France, Fenella G.

frfr@loc.gov

FAIC, Library of Congress, USA

Toth, Michael B.

mbt.rbtot@gmail.com

R. B. Toth Associates, USA

Hansen, Eric F.

ehan@loc.gov

FAIC, Library of Congress, USA

The Library of Congress has expanded its digital spectral imaging research of humanities artefacts that reflect the history of the United States, with the development of advanced imaging techniques that provide data for the studies of manuscripts that span the centuries: Portolan Charts – from 1320-1633, Jefferson's handwritten draft of the Declaration of Independence from 1776, and Herblock's political cartoons from 1929-2001. Using standardized digital imaging techniques, the Library of Congress Preservation Directorate is providing preservation scientists, conservators and humanities scholars with access to digital information on historic and fragile documents with conservation-safe, non-destructive technologies. This provides data for greater understanding of the original object, including revealing creation techniques, and identifying the origin of the substrate (paper, parchment) and media. The Library plans to host this data in standardized format for access as part of a broader preservation database of scientific reference materials of naturally-aged substrate, media (inks, colorants and pigments), treatment effects, environmental data and other document production and creation information. These recent advances in technology and digital access have paved the way for the improved utilization and interpretation of scientific analyses to contribute to scholarly interpretations of heritage materials.

1. Integration of Imaging and Data Management for Discovery

The Library of Congress has implemented digital spectral imaging of the following objects to collect

preservation data, scholarly information and a cross-section of data on cultural heritage old and new:

- Portolan Navigational Charts:¹ Imaging is being used to non-destructively characterize a range of pigments, details of compass points, creation techniques and tools, and potential palimpsest information.
- The handwritten draft of the Declaration of Independence:² Hyperspectral imaging revealed layers of changes and different inks, and offered new insights into the original Jefferson text that was crossed out and overwritten.
- Herblock Political Cartoons:³ A selection of the large original drawings were spectrally imaged to assess the condition of light sensitive inks, also revealing details of the drawings not previously discovered.

The application of digital hyperspectral imaging and associated non-destructive technical analyses to key cultural objects at the Library of Congress required the integration of complementary data to address a wide range of preservation and scholarly challenges. The advanced imaging data is incorporated into ongoing humanities studies of objects, generating digitally linked data sets in standardized format for Internet access. The non-destructive imaging capabilities allow researchers to characterize pigments and media on the artefact, retrieve hidden and lost text, and illuminate production methods of a range of cultural objects. Characterization of a range of materials has been enhanced through the development of a standardized spectral reference sample set, virtually eliminating the need for any sampling. Integration of the data from these technological advances with information from other preservation studies, allows greater scholarly access to the information available from fragile historic documents on parchment and paper.

Application of non-destructive imaging techniques allows the equivalent of optical archaeology of these manuscripts and documents. The profound advantage of this technique is to provide a wealth of information and data – while minimizing and preventing further deterioration of fragile heritage items through handling and invasive analytical techniques. Integrating the range and volume of data collected from any one of these objects in a cohesive data set requires the development of a spatial "map" of the object with layers of information that relate to specific points and details on the document. This range of data can include materials characterization of pigments, colorants and other components, scholarly interpretations of text, organic and inorganic compound information, topographical layering of additions to the document, and evidence about equipment and tools used in

the creation of the document. All this information adds to the interpretation of cultural objects and advances the role of non-destructive heritage science techniques in humanities research.

2. Imaging

The Preservation Research and Testing Division at the Library has developed its MegaVision-Equipoise Spectral Imaging System for preservation research into a range of United States' Top Treasures and international objects of cultural import from across the vast Library collection of nearly 145 million items, including: More than 32 million cataloged books and other print materials in 470 languages; over 62 million manuscripts; the largest rare book collection in North America; and the world's largest collection of legal materials, films, maps, sheet music and sound recordings. Collecting accurate standardized imaging data with this or any other imaging system requires integration and management of three critical factors: 1) the imaging system, 2) standardized metadata capture, and 3) efficient work processes.

The Megavision-Equipoise imaging system has been customized for cultural studies, including a focus on conservation safe lighting that minimizes light on the document while generating a high resolution image. The system collects a "cube" of standard registered images from the ultraviolet through the visible spectrum to the infrared (approximately 365nm – 1050nm) with a MegaVision 39 Megapixel monochrome camera and Equipoise LED EurekaLights. Integrated side-light panels add to the previous lighting system. The inclusion of a hyperspectral scientific reference sample collection has greatly aided the characterization of a range of materials. Standardized metadata with imaging and illumination information, as well as information about the document and its content is captured with the PhotoShoot™ software, which embeds metadata in the header of the image files, based upon EXIF and IPTC standards. The quality of the spectral imaging data relies on the adaptation of proven work processes to the requirements of each object. New imaging applications can then focus on acquiring scholarly information.⁴ The process flow continues beyond the actual imaging, to include image analysis, potential post processing and validation of the data – an iterative loop requiring input from a range of personnel and expertise – requiring additional time, resources, and management.

3. Data Management

For the collection, coordination, access and presentation of this data, the Preservation Directorate is developing a comprehensive approach for access to digital files in a universally accessible

format. This format will utilize an RFD framework for international collaboration and ease of access, a critical component being standardization of file formats from a range of instrumentation. The approach is to integrate scientific and intellectual scholarly information, including the interpretive data required for humanities researchers to utilize the information effectively. For example, this interpretive data could involve identification of a pigment that was not discovered until after the time period attributed to the document, or lost text that changed the interpretation of the document, and provided greater insight into the thought process of the creators. The range of data is being structured in a comprehensive format with the utilization of "scriptospatial" digital objects – essentially a geospatial information system (GIS) for documents. This data organization includes addressing the challenge of presented and linking data across a collection of items to show the scientific visualization and representation of changes across geographical locations, chronological time periods, changing use of materials, and the development of new production techniques.

4. Conclusion

These hyperspectral imaging studies from the 1400s to the twenty-first century are revealing a range of scholarly and preservation information about seminal objects that represent specific aspects of their era. Linking non-destructive scientific imaging and digital technologies with humanities research augments the preservation of these often fragile cultural artefacts, while improving and increasing access for researchers, with extensive intellectual implications. This has created a powerful tool for probing the past; revealing levels of data and raising and answering further questions about the documents, questions that could not previously be contemplated due to missing and incomplete information. The combination of scientific and humanities research allows an enhanced dialogue between researchers, harnessing the strengths of researchers and scholars in each field, and creating a more effective interpretation of data.

Often the focus and import of this exchange of information and newly acquired data is on the movement of information from scientific analyses to the humanities. However it should be noted that the flow of data is not one-way. Humanities researchers and scholars provide knowledge of past eras and culturally related information that can prove of great benefit in the interpretation of scientific analyses, such as the knowledge of local and or cultural practices, and treatises on commonly used materials and practices. The comprehensive presentation of this data in a

form that allows these two complementary streams of research to be linked and integrated greatly enhances this dialogue. This ongoing and iterative dialogue is a critical component for advancing and preserving cultural knowledge throughout the centuries – not only from the past, but also through preservation research into modern fugitive materials and media. This collaborative research can only be accomplished with the ability to capture standardized images, data and metadata from scientists and scholars and integrate them into a common data set, advancing the integration of heritage science and humanities research. The combination of technological advances and structured data access enhances accessibility to original scientific data files and images. Interpretation of this data is enhanced by ease of access to integrated data files. For humanities research, this provides access to linked data files, increasing the intellectual capacity to harness and share knowledge internationally through digital and technology advances.

References

- Bouissac, P.** (2006). 'Probing Pre-historic cultures: data, dates and narratives'. *Rock Art Research*. **Vol. 23**: 89-96.
- Casini, A, et al.** (1999). 'Image spectroscopy mapping technique for non-invasive analysis of paintings'. *Studies in Conservation*. **Vol. 44**: 39-48.
- Christens-Barry, W., Boydston, K., France, F.G., Knox, K., Easton, R.L. and Toth, M.B.** (2009). 'Camera system for multispectral imaging of documents: Digital Imaging Sensors and Applications'. *IS&T/SPIE 21st Annual Symposium Electronic Imaging, Science and Technology*. **Vol. 7249**: 1-10.
- Ciula, A., Spence, P. and Viera, J.M.** (2008). 'Expressing complex associations in medieval historic documents: the Henry III fine rolls project'. *Literary and Linguistic Computing*. **Vol. 23, No. 3**: 311-325.
- Dawes, S.S.** (2009). 'Governance in the digital age: A research and action framework for an uncertain future'. *Government Information Quarterly*. **Vol. 26**: 257-264.
- Emery, D., France, F.G. and Toth, M.B.** (2009). 'Management of spectral imaging archives for scientific preservation studies'. *Archiving 2009: Preservation Strategies and Imaging Technologies for Cultural Heritage Institutions and Memory Organizations*. VA: Society for Imaging Science and Technology, pp. 137-141.
- France, F.G.** (2007). 'Managing digital image repositories as key tools in the preservation of cultural objects'. *Imaging Science and Technology Conference*. Arlington, pp. 117-121.
- 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*. **Vol. 80, No. 1**: 33-59.
- Grenacher, F.** (1970). 'The Woodcut Map: A form-cutter of maps wanders through Europe in the first quarter of the sixteenth century'. *Imago Mundi*. **Vol. 24**: 31-41.
- Jessop, M.** (2008). 'Digital visualization as a scholarly activity'. *Literary and Linguistic Computing*. **Vol. 23, No. 3**: 281-293.
- Jessop, M.** (2008). 'The inhibition of geographical information in digital humanities scholarship'. *Literary and Linguistic Computing*. **Vol. 23, No. 1**: 39-50.
- Nayar, S.K., Branzoi, V. and Boult, T.E.** (2006). 'Programmable imaging: Towards a flexible camera'. *International Journal of Computer Vision*. **Vol. 70(1)** : 7-22.
- Sculley, D. and Pasanek, B.M.** (2009). 'Meaning and mining: the impact of implicit assumptions in data mining for the humanities'. *Literary and Linguistic Computing*. **Vol. 23, No. 4**: 409-424.

Notes

1. Portolan charts are early nautical navigational maps based on realistic descriptions of harbours and coasts. These were first made in the 1300s in Italy, Portugal and Spain (*portolan* comes from Italian, *portolano*, meaning "related to ports or harbours"). The charts cover the period from 1320 to 1633, and were created on vellum or parchment.
2. The draft of the Declaration of Independence was handwritten in 1776 by Thomas Jefferson in iron gall ink on laid paper, with corrections and changes by Benjamin Franklin and John Adams.
3. Created by Herbert Block from 1929 to 2001, these cartoons represented the penmanship of a man who influenced public opinion in America throughout his 72-year career.
4. As new imaging applications are integrated into the process, rigorous attention to imaging details allows an efficient capture of data with careful document handling, standardized imaging procedures (including lighting, relative humidity and temperature control) and data management. These factors are supported with environmental management, security, contingency planning and IT infrastructure.

Building Dynamic Image Collections from Internet

Fu, Liuliu

luna.foe@gmail.com

Old Dominion University, USA

Maly, Kurt

Old Dominion University, USA

Wu, Harris

hwu@odu.edu

Old Dominion University, USA

Zubair, Mohammad

Old Dominion University, USA

People often want to collect and utilize free, publicly available images on a given subject. Image sharing systems such as Flickr store billions of user-contributed images. While such systems are designed to encourage user contributions and sharing, they are not well-organized collections on any given subject. We propose an approach that systematically harvest images from Internet and organize the images into an evolving faceted classification. We implemented a prototype to continuously harvest the most popular images on Flickr related to African American history, and organize them into an evolving faceted classification collaboratively maintained by users. The same approach can be applied to other digital humanities resources on the Internet. The talk will elaborate the details of technical design and prototype implementation, and discuss evaluation results.

2. Introduction

Flickr hosted over 4 billion images as of October 2009 and is growing by about 4 million pictures a day. Facebook hosted 15 billion photos and Imageshack hosted 20 billion images as of April 2009. Other photo sharing sites such as Picasa, Multiply and PhotoBucket also host billions of images [Schonfeld, 2009]. In contrast, organized public domain image collections are relatively scarce [Wikipedia: 'Public domain image resources']. The largest public domain image repository, Wikimedia Commons, reached 5 million images as of September 2009.

It would greatly benefit the humanities community if images in those image sharing sites can be organized and utilized. We propose an approach that systematically searches and harvests images (actually the link to the image and metadata, but not image itself) from image sharing sites,

and organizes the images into a multi-faceted classification. The data harvesting is performed on a continual basis, and the classification evolves over time. Besides automated programs, the approach utilizes collaborative human efforts to improve the quality of collection. We implemented a prototype that builds a dynamic image collection on African American History from the most popular images on this subject on Flickr.com. Our fundamental belief is that a large, diverse group of people (students, teachers, etc.) can do better than a small team of librarians or editors in constructing a multimedia collection at virtually no cost.

Note that not all the images on those sharing sites are copyright-free or have a creative commons license. However, most of the sites allow other websites to directly link to their images if the images are marked as public access by their contributors, and if credits are properly given. Our approach displays images through embedded image URLs but does not download the images from their original sources.

3. Related Work

Many are trying to utilize the images in fast-growing photo sharing and social networking sites. For example Getty Images, the leader in stock photography, hires image editors to select most popular Flickr images and obtain copyright from individual contributors, then sells the images for \$5 per image (<http://www.gettyimages.com>). Computer-graphics researchers at the University of Washington have utilized Web images to digitally reconstruct buildings in 3-D. For example, based on 150,000 publicly accessible Flickr pictures of Rome, the program automatically re-created the Colosseum, Trevi Fountain, and the outside and inside of St. Peter's Basilica, among other Roman icons. The technique can be used to make virtual-reality experiences for tourism, auto-build cities for video games and movies, or help digitally preserve and study historic cities that are being destroyed by human-caused or environmental factors [Jaggard 2009].

Researchers have argued for building an academic Flickr, or an academic photo sharing site in general: a net-based service that would enable faculty and researchers to post and share images with scholarly value, either with the general community, or pursuant to any associated rights, to restricted-use populations [P. Brantley's blog]. For example, a group at Lewis & Clark College in Portland is in the process of developing an educational collection of contemporary ceramics images using the photo sharing site Flickr [McWilliams 2008].

Our project attempts to build free, well-organized topical images collections from the images

contributed by Internet users, to support education or research objectives. While most photo sharing systems support keyword-based search utilizing user-contributed metadata, none of them support browsable hierarchies that allow users to explore a given subject in depth. Using librarians or images editors to manually construct a topical collection is cost prohibitive, and unfeasible if the collection needs to keep up with rapidly growing data sources. Our collection-construction approach combines the collaborative concepts of wiki and social tagging systems with automated classification techniques. Our system allows users to collaboratively build a classification schema with multi-faceted categories, and to classify documents into this schema. Utilizing users' manual efforts as training data, the system's automated techniques build a faceted classification that evolves with the growing collection, the expanding user base, and the shifting user interests.

4. Architecture and Prototype Implementation

Our collection construction approach is summarized in Figure 1. The system first collects images (links and metadata such as tags) on a given topic using keyword search, utilizing the APIs (Application Programming Interface) provided by image sharing sites or search engines. For the initial collection, a group of experts or administrators create the initial classification schema and classify a set of images into the initial schema. Utilizing experts' classifications as training data, and also Wordnet and Wikipedia as knowledge bases, the system employs automated techniques (heuristic matching rules and support vector machine-based classifiers) to classify images into the classification schema. In a wiki fashion, users of the image collection can modify and improve the classification schema, and manually classify items into the schema. Users can also assign additional tag or annotations to image objects. Utilizing the additional metadata from users' tagging and annotation efforts and by analyzing users' classification/usage history, the system refines both the classification schema and the item-category associations. The system continues to collect and classify images to stay up-to-date with external image sources.

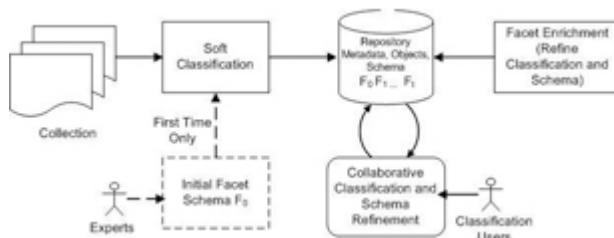


Figure 1. Systematic approach of constructing a topical collection using Internet images.

We built a prototype to construct an image collection on African American History from Flickr. By querying "African American History" in the search field, we extracted metadata for all the images in the result pages: title, url, description, tags, and the contributor. The initial collection contained about 11,000 Flickr images on African American History. Over 3 months the collection has grown to contain about 13,000 images. During the conference we will elaborate the details of technical design, prototype implementation, and the evaluation results. Figure 2 shows the browsing and classification interfaces of our prototype.



Figure 2. Browsing and classification interface of the prototype.

5. Discussion

Evaluation of the prototype in a classroom environment shows promise. Measured by metrics such as precision, recall and image quality (popularity), the prototype is more effective than Flickr in supporting several image retrieval tasks. The evolution of classification shows improvements,

based on user ratings of categories and category-item associations. We conducted interviews and usage observations, which help understand the level of efforts that users spend on tagging and classification. For future research, we are interested in whether social tagging and tag convergence [Muller *et al.* 2008] can be utilized to assist or substitute classification efforts.

Our approach can be applied to other digital humanity resources. For example, we have developed another prototype to construct a dynamic collection of news items on a given topic based on Google News. We believe that a combination of collaborative and automated classification techniques can construct valuable digital humanities collections at low costs.

As far as we know, no one has combined user efforts and automated techniques to build a faceted classification. Several research projects are related to social tagging and classification, however. Several projects attempt to construct tag hierarchies or ontologies, or otherwise harvest the intelligence stored in tags [Heymann and Garcaí-Molina 2006, Schmitz and Patrick 2006, Harris *et al.* 2006]. Our earlier work [Arnaout *et al.* 2008] on faceted classification was presented in the Digital Humanities 2008 conference.

6. Acknowledgements

This project is supported in part by the United States National Science Foundation, Award No. 0713290.

References

- Schonfeld, Erick** (2009). 'Who Has The Most Photos Of Them All?'. *TechCrunch*. April 7, 2009. <http://www.techcrunch.com/2009/04/07/who-has-the-most-photos-of-them-all-hint-it-is-not-facebook/>.
- Wikipedia:Public Domain Image Resources.** http://en.wikipedia.org/wiki/Public_domain_image_resources.
- Jaggard, Victoria** (2009). 'Flickr Pictures Help Build 3-D Rome in a Day'. *National Geographic News*. September 24, 2009.
- Jaggard, Victoria** (September 24, 2009). *Flickr Pictures Help Build 3-D Rome in a Day*. *National Geographic News*. <http://news.nationalgeographic.com/news/2009/09/090924-flickr-rome-build-day.html>.
- Brantley, Peter.** *Design Beyond the Interface* Blog http://blogs.lib.berkeley.edu/shimenawa.php/2008/04/17/ah_screw_the_interface.
- McWilliams, Jeremy** (2008). 'Developing an Academic Image Collection with Flickr'. *Code{4}lib Journal*. 3.
- Muller, M.J., Dugan, C., Millen, D.R.** (2008). 'Metrics for sensemaking in enterprise tag management'. *CHI 2008 Sensemaking Workshop*. Florence, Italy, April 05 - 10, 2008, pp. 1493-1496.
- Heymann, P., Garcia-Molina, H.** (2006). *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. University of Southampton: Stanford Technical Report InfoLab. <http://ilpubs.stanford.edu:8090/775>.
- Schmitz, Patrick** (2006). 'Inducing Ontology from Flickr Tags'. *World Wide Web Conference 2006 (WWW2006)*. Edinburgh, UK., May 22–26, 2006. Collaborative Web Tagging Workshop. .
- Wu, Harris, Maly, Kurt, Zubair, Mohammad** (2006). 'Harvesting social knowledge from folksonomies'. *Hypertext 2006 – Seventeenth ACM Conference on Hypertext and Hypermedia*. Odense, Denmark, 23-25 August 2006.
- Arnaout, Georges, Maly, Kurt, Mektesheva, Milena, Wu, Harris, Zubair, Mohammad** (2008). 'Exploring Historical Image Collections with Collaborative Faceted Classification'. *Digital Humanities 2008*. Oulu, Finland, June 25-29, 2008.

GIS, Texts and Images: New approaches to landscape appreciation in the Lake District

Gregory, Ian

I.Gregory@lancaster.ac.uk
Lancaster University

The use of GIS in historical research, Historical GIS, is now a well established part of the discipline of history. The field has evolved to an extent where it can be shown to have made a significant impact in delivering high-quality research in books and peer reviewed journals including the *British Medical Journal*, *Annals of the Association of American Geographers*, the *American Historical Review*, *Journal of Economic History*, and the *Agricultural History Review*. Most of these studies are, however, largely concerned with quantitative, social science-based approaches to historical research. This paper explores how approaches based on other sources such as texts and images can be used to allow GIS to be applied across the disciplines of the humanities. Early research is already suggesting that it can and indeed a new field, spatial humanities, is increasingly being recognised. This paper will explore one example of this approach focusing on how we can use GIS techniques to integrate historical texts and modern 'born-digital' photographs to gain a better understanding of landscape appreciation in the Lake District.

The paper starts by looking at two early tours of the Lake District, Thomas Gray's proto-Picturesque tour of 1769 and Samuel Taylor Coleridge's 'circumcursion' of 1802. We are currently working to extend this to include a subset of William Wordsworth's work. This project extracted place-names from these texts and matched them to a gazetteer to turn them into GIS form. The advantage of this approach is that once the GIS database has been created the spatial information in the texts can be mapped, re-mapped, queried, integrated with other material, and manipulated in a wide range of ways. The project produced a range of maps including: simple dot-maps of places mentioned, density smoothed maps that use techniques pioneered in epidemiology and crime mapping to summarise complex point patterns, and maps of emotional response to the landscape. Some of these were of the individual texts, some compared and contrasted the different texts. Other forms of analysis integrated data from other sources such as

a Digital Elevation Model of the Lake District, and contemporary population densities. From these we were able to show that Gray followed the main valleys of the Lake District and stayed in towns overnight. He rarely travelled to heights of more than a few hundred feet although the higher peaks, those of over 2,500 feet, attract considerable attention in his writing. Coleridge, by contrast, avoided the populous parts of the Lake District, staying in the Western Fells and climbing Sca Fell, the highest mountain in England, among other things. While his ascent (and hair-raising descent) of Sca Fell is well known, what is more interesting is that much of his account is also concerned with time spent in low places, similar to Gray, and also that he names places of all heights, especially those between 1,000 and 2,000 feet which Gray almost completely ignores. The two tours barely overlap, the only place where they do significantly is Keswick, where Coleridge lived and Gray spent several nights, and the road over Dunmail Raise between Grasmere and Keswick although neither account has much to say about this part of their journey.

This approach takes us into what F. Moretti (2005) has termed 'distant reading,' a methodology that stresses summarising large bodies of text rather than focusing on a few texts in detail. We also wanted to explore whether GIS could help with more traditional approaches to reading. To this end we created a KML version of the GIS implemented in Google Earth. This placed a text on the bottom half of the screen with a Google Earth map on the top-half. Superimposed on the map were the locations mentioned in the texts, which can be switched on and off in various ways, and a contemporary map showing the Lakes in 1815. This allows the reader to read the text while following the locations named using either Google Earth's modern arial photographs or the historical map as a backdrop. This therefore enriches the experience of close reading of the text by visualising and contextualising the places mentioned. Given the numbers of places mentioned by both authors even someone highly familiar with the Lake District is unlikely to be able to accurately locate all of these mentally. An alternative approach that this framework provides is for the user to click on a location and ask "what have the different writers said about this place?" To enrich this further we allow users to link from the site to the photographic website Flickr. Flickr allows people to upload and share their digital photographs. Users can tag these with metadata such as 'landscape' or 'mountain' and can also add 'geo-tags' a latitude and longitude that give the photo a location. Using these allows us to link from our texts to allow the user to see what people have photographed nearby.

The initial idea behind linking to Flickr was simply to demonstrate what the different areas of the Lake District looked like to an audience who might be unfamiliar with it, and thus to assist the in-depth reading. It became apparent however that there are pronounced geographies within Flickr – some areas are extensively photographed and some ignored, while the different tags that people place on their images also have pronounced geographies. As Wordsworth is claimed to have extensively influenced the way people today view the landscape, particularly in the Lakes, which poses the question "is there are relationship between the geographies of Wordsworth's writing and the geographies of Flickr." Using the Flickr API we were able to extract the number of photographs geo-tagged to locations in cells of approximately 1km square across the whole of England. This could be done for all photos or those with specific tags such as 'mountain(s)' or 'tree(s).'!

Mapping all photographs produces some interesting geographies, in particular, most photos seem to be taken in the urban centres or the main valleys. Minor roads such as that over the passes of Wrynose and Hardknott, also seem to encourage photography. It may be therefore that modern visitors to the Lake District, at least as represented by people who upload geo-tagged photographs to Flickr, follow a tour that is more like the Picturesque tours of Gray than the Romantic experiences of Coleridge or Wordsworth. In this way we are able to return to distant reading and to integrate two apparently incompatible sources: historical writings and modern digital photographs.

This paper thus demonstrates the potential of using geo-spatial approaches to integrate disparate and apparently incompatible sources. In it we have integrated historical texts, historical maps, modern environmental information giving information on the topography, statistical data from the census giving population densities, and born digital images from Flickr. By bringing them together we have been able to shed new light on a specific topic, landscape appreciation in the Lake District. The implications, however, are far broader. The amount of geo-referenced data available from multiple sources is increasing exponentially. This can be expected to continue particularly given the growth of user-generated content and the availability of techniques to automatically geo-reference texts. The challenge is to use these sources in innovative ways to shed new insights into research questions in the humanities. If this can be done successfully it will lead to a re-awakening of the importance of geography to the humanities.

Capturing Visitor Experiences for Study and Preservation

Georgina Guy

georgina.guy@kcl.ac.uk
King's College London

Stuart Dunn

stuart.dunn@kcl.ac.uk
King's College London

Nicolas Gold

nicolas.gold@kcl.ac.uk
King's College London

The Courtauld Gallery occupies a unique position in central London as a university art museum and key facilitator for collaborative opportunities between research academics and curatorial practitioners. Given the context of the gallery as an institution housed within a building not purpose-built for the function of exhibition, the ways in which the gallery space acts as a directive on visitors' viewing patterns is of particular interest. This paper introduces a methodology of documenting and visualizing those patterns.

The research is approached from the perspective of a performance specialist (Guy), investigating the curated gallery as a place for performance and the visitors' role within such exhibition spaces. This has led to the generation of questions vital to new methodological approaches for exploring data only existent in the form of events, and to future processes of gallery operation and evaluation.

This paper focuses on the empirical aspects of this work, reporting on a project to create digital objects, displayable in Virtual World platforms such as Google Earth, from the experience of visitors to the Courtauld Gallery, London. The digital objects consist of visualisations in virtual space and time of visitor experiences, documented in KML, and represented using the Google Earth platform.

1. Studying Visitor Experiences

To date, the evaluation of exhibitions has largely been based on attendance numbers, with very little attention given to actual visitor behaviours within gallery environments. Traditionally, where visitor behaviours have been observed, methods for achieving this have been based on pen and paper recording using methods such as those developed by Space Syntax (Space Syntax, 2010). Capturing

this data in digital form will allow a more thorough and formal analysis of the ‘success’ of exhibitions by permitting the replay in both time and virtual space of visitors’ behaviours and their interaction with staff, other visitors, and gallery exhibition materials. More specifically, we are concerned with:

- patterns of movement within gallery spaces
- specific pathways constructed by individual visitors through the museum
- duration of engagement with individual exhibits
- actions and interactions of gallery visitors

This information makes possible an analysis of how visitors explore exhibitions that is not preconceived but observed. Anonymous representations of visitor behaviours can be offered back to the gallery prompting curatorial assumptions to be validated or, where appropriate, reconsidered in light of evidential data about real patterns of visitor behaviour. This, in turn, can have important implications for maximizing public engagement within exhibition contexts and ensuring efficiency of interaction between staff and visitors, as well as other aspects of social and economic exchange. Using geovisualization techniques to generate the interactive maps based on individual experience raises questions about how possible it is to produce and manage the documentation of human behaviour.

2. Technological and Methodological Issues

Ideally, visitors would be totally oblivious to the data capture process or at least, such technological means as are necessary for capture would be non-intrusive. The demands of the research problem require high fidelity of location, orientation, and behaviour making the technological issues more complex than simply determining approximate location. Borriello et al. (2005) report that GPS systems do not work reliably inside buildings, wifi systems require calibration and achieve accuracy of only about 3 metres, and others require infrastructure installation. This is rarely possible in protected buildings such as the Courtauld, especially on a temporary basis. Consequently, the method described below was developed (related approaches will be discussed in presentation).

A pilot study was carried out prior to the main observation period in the Courtauld’s Frank Auerbach exhibition (Courtauld Gallery, 2009).

3. Data Capture Method

The method used is predominantly based on field observation in two forms. These approaches are

influenced by the observation methods used by Space Syntax, a company with an established history of evidence-based design and evaluation for buildings and urban spaces. Space Syntax observations are undertaken using pen and paper.

The first method involves a human observer tracking the movement pattern of visitors around parts of the gallery in order to observe particularized routes specific to individual visitors. This is facilitated by the use of Tablet PCs and custom-developed software which displays an editable floor plan of the gallery, divided into map tiles based on the gallery rooms. In the case of the Courtauld, the map tiles are 610x365 pixels. The actual room size in the pilot is approximately 1290.5cm x 772cm = 12.9x7.7m. By moving the digital pen around the image the pathway of an individual can be recorded on the map and the movement documented with an x,y pixel reference and a timestamp. Duration and location are recorded both when the visitor is moving and stationary. Additional coding concerning, for example, the activity undertaken at any given moment, such looking at a map or signage or taking a photograph, can also be coded against points on each pathway.

The second method involves participants being asked to complete an exit questionnaire detailing their familiarity with the specific gallery and exhibits as well as more general experience of exhibition environments. This information is cross-referenced to the trace of each visitor’s pathways.

A screenshot from the software can be seen in Figure 1, showing a gallery floorplan with a partial trace. Figure 2 shows the simple comma separated values (CSV) data produced by the trace activity in realtime (the fields being *X*, *Y*, *timestamp*, *visitor activity*, and *map file used*). The system defaults to tracking movement as the pen is drawn across the screen, however, the observer can simply switch the “mode” of the current point by selecting one of the buttons to the right. Where data about interaction with an exhibit is necessary, the point of the observee’s interest can be noted by simply pointing at it on the diagram after selecting either “Sign” or “Exhibit” as the object of interest. The final system was developed in Borland Turbo Delphi with early prototyping undertaken in Processing.

For the Courtauld case study, the maps are scaled such that 1 pixel is approximately 2cm² of real gallery space. The maps are converted from architectural plans and the location and size of furniture in the gallery included from measurement in the galleries themselves. Visitor location is recorded as a relative pixel position from the top left corner of the image. After capture this data can be transformed to generate real visitor positions in the room for subsequent visualisation.

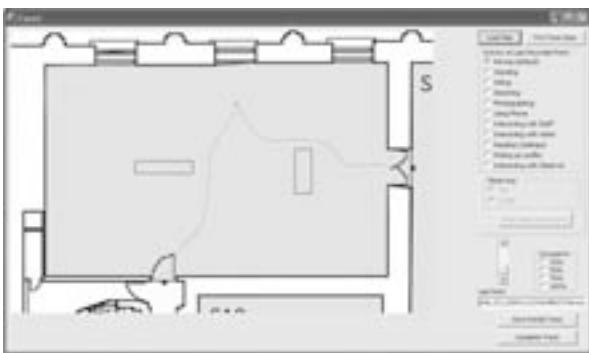


Figure 1: Screenshot from Data Capture Application

F2g15.bmp.partial.20091028093547421.csv - Notepad	
File	Edit
Map Title Changed From:	
Map Title Changed To: C:\temp\f2g15.bmp	
START OF TRACE	
270, 427, 20091028093607750, MovIngr, C:\Temp\f2g15.bmp	
270, 426, 20091028093608546, MovIngr, C:\Temp\f2g15.bmp	
270, 425, 20091028093608078, MovIngr, C:\Temp\f2g15.bmp	
270, 424, 20091028093608509, MovIngr, C:\Temp\f2g15.bmp	
270, 423, 20091028093608136, MovIngr, C:\Temp\f2g15.bmp	
270, 422, 20091028093608218, MovIngr, C:\Temp\f2g15.bmp	
270, 421, 20091028093608263, MovIngr, C:\Temp\f2g15.bmp	
271, 420, 20091028093608298, MovIngr, C:\Temp\f2g15.bmp	
271, 419, 20091028093608359, MovIngr, C:\Temp\f2g15.bmp	
,	
478, 248, 20091028093620000, MovIngr, C:\Temp\f2g15.bmp	
477, 248, 20091028093620062, MovIngr, C:\Temp\f2g15.bmp	
478, 247, 20091028093620109, MovIngr, C:\Temp\f2g15.bmp	
480, 247, 20091028093620140, MovIngr, C:\Temp\f2g15.bmp	
481, 246, 20091028093620171, MovIngr, C:\Temp\f2g15.bmp	
482, 246, 20091028093620281, MovIngr, C:\Temp\f2g15.bmp	
483, 243, 20091028093630312, MovIngr, C:\Temp\f2g15.bmp	
484, 243, 20091028093630328, MovIngr, C:\Temp\f2g15.bmp	
484, 246, 20091028093630437, Standing, C:\Temp\f2g15.bmp	
480, 63, 20091028093632311, OBSERVING, C:\Temp\f2g15.bmp	
,	
484, 246, 20091028093620437, Standing, C:\Temp\f2g15.bmp	
483, 242, 20091028093635078, MovIngr, C:\Temp\f2g15.bmp	
483, 241, 20091028093635125, MovIngr, C:\Temp\f2g15.bmp	
483, 239, 20091028093635175, MovIngr, C:\Temp\f2g15.bmp	
483, 237, 20091028093635187, MovIngr, C:\Temp\f2g15.bmp	
482, 236, 20091028093635203, MovIngr, C:\Temp\f2g15.bmp	

Figure 2: Extract from CSV data file

3.1. Visualisation of Results

The relative positional information in the CSV file(s) is converted to absolute positions and stored in KML. The timestamp stream that indicates where a visitor was at a given time can be used to generate a trace of movement (and speed).

For the purposes of the visualization, the floorspace of the Auerbach, as defined by the map tiles in the data capture software, was rotated from its SE-NW alignment (see Figure 3). This was necessary to facilitate the georeferencing, or conversion of the x,y pixel data points into decimal degree coordinates. It was therefore possible to make the observation that the gallery room equated to 0.000118×0.000027 decimal degrees. The pixel readings (r) can therefore be converted to decimal degrees (d) using the simple conversion formula $d = W.\text{longitude} - (0.000118/610)*r$. For the latitude readings, this is repeated, but substituting $(0.000027/325)$. An arbitrary altitude value, starting at 0.001 at the first reading, and increasing sequentially by 0.001 throughout the CSV dataset was added to the

datapoints. In the visualization (see Figure 4), this gives the impression of the pathway rising as the visitor progresses through the gallery and through time: in this, it follows the principle of the ‘space time cube’, developed elsewhere (Kraak, 2008). As more data is added, this will allow us to build complex structured visualizations, which will add significant support to interpreting the visitors’ uses of, and interactions with, the space. These will be reported in full in the full presentation of this paper.

3.2. Conclusion

At the time of writing, the technology platforms are fully developed, ready for use and have been tested (with subsequent enhancement) in a pilot session in the gallery. The main observations will take place in November and December 2009 during the Courtauld Gallery’s exhibition of *Frank Auerbach: London Building Sites 1952–62*.

This paper has presented an approach to capturing visitor experience and interactions in a gallery using methods based on tried and tested approaches, but augmented by digital tools. The data generated can subsequently be visualised in virtual space and time to allow questions of performance practice to be addressed.



Figure 3: The Auerbach gallery in Google Earth, and the area of the map tile on which the trace was taken, aligned on a N-S bearing. (© 2009 Google, Image © 2009 Bluesky)

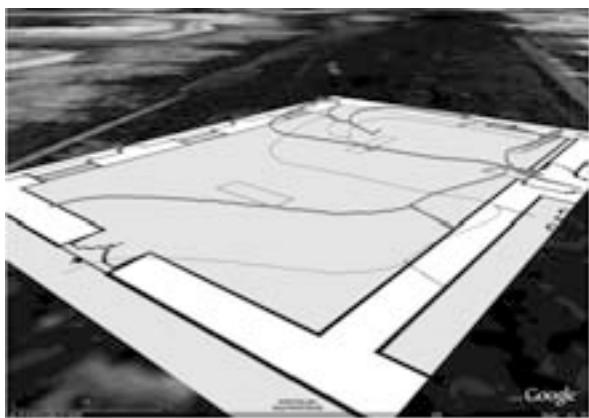


Figure 4. Representation of visitor pathway from the pilot study in 'space time cube' format, represented in Google Earth. (© 2008 Google, Map Data © 2009 Tele Atlas, Image © 2009 Bluesky)

References

- Boriello, G., Chalmers, M., LaMarca, A., Nixon, P. (2005). 'Delivering Real-World Ubiquitous Location Systems'. *Communications of the ACM*. **48(3)**: 36-41.
- Courtauld Gallery. 2009. <http://www.courtauld.ac.uk/gallery/exhibitions/2009/auerbach/index.shtml> (accessed 3rd March 2010).
- Kraak, M.-J. (2008). 'Geovisualization and Time - New Opportunities for the Space-Time Cube'. *Geographic Visualization: Concepts, Tools and Applications*. Dodge, M., McDerby, M., Turner, M. (eds.). London: Wiley.
- Space Syntax. 2010. <http://www.spacesyntax.com> (accessed 3rd March 2010).

The Diary of a Public Man: A Case Study in Traditional and Non-Traditional Authorship Attribution

Holmes, David I.

dholmes@tcnj.edu

The College of New Jersey, USA

Crofts, Daniel W.

crofts@tcnj.edu

The College of New Jersey, USA

In 1879 the *North American Review* published in four separate monthly installments excerpts from "The Diary of a Public Man" in which the name of the diarist was withheld. It was, or purported to be, a diary kept during the "secession winter" of 1860-61. It appeared to offer verbatim accounts of behind-the-scenes discussions at the very highest levels during the greatest crisis the US had ever faced. Interest in this real or purported diary was considerable. The diarist had access to a wide spectrum of key officials, from the South as well as the North, gave a number of striking anecdotes about Abraham Lincoln, and provided an important account of events at Washington during the critical days just before the Civil War.

A detailed study of the Diary was conducted by Frank Anderson in 1948 in his book *The Mystery of "A Public Man"*. Anderson argues that the Diary is part genuine and part fictitious with two of the three striking Lincoln incidents appearing to be inventions, along with other so-called "interviews" with prominent figures. He believes that, as a core, there is a genuine diary kept by Samuel Ward (1814-1884) at Washington during that winter, and that it is possible that the editor of *North American Review*, Allen Thorndike Rice, may have assisted in the process of embellishment. William Hurlbert (1827-1895), he argues, may also be involved, since the style of the Diary has a good deal of Hurlbert's pungency. Others have suggested that the diarist might be Henry Adams (1838-1918), who enjoyed close access to William Henry Seward who became Lincoln's Secretary of State and was a central figure in the Diary. Certainly the fact that, over a century after its publication, the authorship has remained undetermined is proof that the work of all those who may have shared in its preparation and publication was cleverly done.

1. Traditional Attribution

This paper argues that the diarist was not Samuel Ward; it was, instead, William Hurlbert. The preponderance of the evidence also suggests that the Diary may well be a legitimate historical document.

The diarist was not simply an observer but very much a participant-observer. One key circumstance would have impeded Ward. At the precise time the Diary was being penned, he was busily engaged in writing a memoir of his experiences in the California gold fields in 1851-52. His recollections were published in a New York weekly, starting on January 22nd 1861, and concluding abruptly on April 23rd 1861. A great deal of internal evidence suggests that the Diary and the Gold Rush memoir could not have been written by the same person, even allowing for their radically different subject matter. Ward's sentences sometimes meander in a Baroque manner, he often alliterates, peppers his narrative with Spanish and French expressions, and has a habit of encasing unusual words or phrases within quotation marks. By contrast the Diary is fast paced and immediate, with a style running towards active verbs accompanied by adverbs of a certain type.

In William Hurlbert, however, we find a newspaper writer whose style had sweep and dash. At the very moment when the biggest story he had ever witnessed burst to attention, he had no job, but nonetheless, had access to a remarkably wide range of prominent people. The Southern-born Hurlbert also had more basis than Ward to have developed close ties with leading Southerners. A comparison of the Diary with things known to have been written by Hurlbert yields some demonstrable parallels, not least in the number of signature words used in the Diary. Some specific features of the Diary also point to Hurlbert rather than Ward, for example twice the diarist mentions Josiah Quincy (1772-1864), the retired President of Harvard, who had been an important influence in Hurlbert's young life. There are circumstances, too, that suggest why Lincoln might initially have encountered Hurlbert and why he might have welcomed a repeat visit.

Concerning its legitimacy, in a number of crucial particulars the Diary conveys an on-the-spot immediacy that would have been almost impossible to recreate even months after the fact, let alone years; for example the unfolding story of Lincoln's secret and circuitous trip to Washington in late February and the diarist's delayed realization that Seward warned Lincoln to undertake it. The diarist expresses repeated concerns about the potential economic effects of secession, concerns which were quickly subordinated once the war started. The diarist also demonstrates an excellent ear in his

accounts of his interviews with others, in particular their personal mannerisms. In all its particulars, the Diary synchronizes perfectly with the way events unfolded at that time.

2. Non-Traditional Attribution

For testing and validating the stylometric techniques involved in this phase of the study, preliminary textual samples were taken from prominent diarists of that era, George Templeton Strong, Gideon Welles, and Salmon Chase. Analysis of the top 50 frequently occurring function words using what is now known as the "Burrows" approach involving principal components analysis and cluster analysis showed clear discrimination between writers and internal consistency within writers. Textual samples were then taken from three candidate authors of the Diary, namely Samuel Ward, Henry Adams and James Harvey, with the "Burrows" approach once again indicating remarkable internal consistency and clear between-writer discrimination.

Four textual samples each of approximately 3,000 words, representing in total about 2/5 of the work, were taken at various places throughout the Diary, being sufficiently spaced to enable a valid check to be made on internal consistency. The Diary samples showed excellent internal consistency, suggesting single authorship which would refute Anderson's "cut and paste" theory. They appeared to be quite distinct from the samples of the writings of Adams, Harvey and Ward.

Focus then moved to the two main contenders for authorship, Ward and Hurlbert. Carefully controlling for genre in the selection of the textual samples from Hurlbert and Ward, subsequent multivariate analyses on high-frequency function words showed discrimination between these two writers, along with internal consistency. For the attributional stage of the research discriminant analysis was employed. The samples from the Diary, Hurlbert and Ward were divided into smaller sizes in order that as many high-frequency function words as possible could be used without violating the assumptions underlying the technique. All 12 Diary samples were placed into the Hurlbert group.

Finally, the "Delta" technique, proposed by Burrows and refined by Hoover, was employed using the 100 most frequently occurring words in the pooled corpus and on four potential authors of the Diary: Ward, Adams, Harvey and Hurlbert. The closest "match" to the Diary using Delta and its variants was indeed Hurlbert.

3. Conclusion

The non-traditional stylometric analysis has supplied objective evidence that supports traditional scholarship regarding the problem of the authorship of the Diary. The likelihood that the entire document was written by one person is very strong. William Hurlbert has been pinpointed, to the exclusion of all others, as the Diary's author. Much of the Diary could never have been concocted after the fact; the chances are that the entire document is authentic.

References

- Anderson, F.M.** (1948). *The Mystery of "A Public Man"; A Historical Detective Story*. Minneapolis: University of Minnesota Press.
- Burrows, J.F.** (1992). 'Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information'. *Literary and Linguistic Computing*. 7: 91-109.
- 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. (ed.)** (1949). *Sam Ward in the Gold Rush*. Stanford: Stanford University Press.
- Hoover, D.L.** (2004b). 'Delta Prime?'. *Literary and Linguistic Computing*. 19: 477-495.

Using the Universal Similarity Metric to Map Correspondences between Witnesses

Holmes, Martin

mholmes@uvic.ca

University of Victoria

Thomas Sonnet de Courval's satirical work, the *Satyre Menippée du mariage*, was initially published in 1608. In 1609, an expanded version appeared, with the addition of a second satire, the *Timethélie, ou Censure des Femmes*. In 1621, the first satire appeared in a new edition titled *Satyre sur les Traverses du Mariage*, then in the following year, Sonnet de Courval published his *Œvres Satyriques*. The *Œvres* includes 12 satires, with the final six consisting of fragmented, re-organized and re-edited versions of the *Satyre Menippée* and the *Timethélie*. A new edition of the 1609 text was published in 1623, edited by the publisher and probably without Courval's consent. In 1627, two more editions of the *Œvres Satyriques* appeared. (Coste 340-341).

The *Mariage sous L'ancien Régime* project has already digitized the 1609 text and most of the 1621, and will be working on other editions in the future. Our objective is to produce a genetic edition of those parts of Courval's work which bear on marriage; in particular, we would like to map the process by which the original two satires (*Menippée* and *Timethélie*) were re-constituted as six satires in the *Œvres Satyriques*. Since the texts are lengthy (the 1609 text runs to nearly 3,000 lines), we have begun to investigate ways to automate this mapping to some degree, and in particular, methods of measuring similarity between two pieces of text. In particular, we needed to find a way to detect corresponding lines between two witnesses, even when those lines might have been both relocated and altered.

The Universal Similarity Metric is a method of measuring the similarity of two sets of data based on Kolmogorov complexity. It is described in Vitanyi (2005) as "so general that it works in every domain: music, text, literature, programs, genomes, executables, natural language determination, equally and simultaneously" (1). A practical implementation of this metric can be achieved using data compression, according to the following formula, where x and y are the two pieces of data being compared:

$$\text{NCD}(x, y) = \frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))}$$

(Vitanyi 2)

NCD is "Normalized Compression Distance", an expression of the similarity of x and y; C(x) and C(y) are the respective lengths of the two compressed inputs; and C(xy) is the length of the compressed concatenation of x and y. The resulting NCD is a value between 0 and 1, where proximity to zero indicates greater similarity. This metric has been widely used in the sciences; for instance, Krasnogor and Pelta (2004) describes its use to measure the similarity of protein structures, and Li et al. (2004) apply it to evolutionary trees and to building language trees based on text corpora. Its universality and simplicity suggested that it might be the ideal tool to discover correspondences between lines and line-groups at different points in two of Courval's texts. To test it, I have created a prototype application using Borland Delphi. These are some example values generated with the prototype (Figure 1):

Text 1	Text 2	NCD Score
These two lines are absolutely identical.	These two lines are absolutely identical.	0,0000000
The vndiscovered country, at whose sight	The vndiscovered Countrey, from whose Borne	0,3673469
To be, or not to be, I there's the point,	To be, or not to be, that is the Question:	0,4545455
To Die, to sleepe, is that all? I all:	Whether 'tis Nobler in the minde to suffer	0,7234043
これは日本語です。	To sleepe, perchance to Dreame; I, there's the rub,	0,8392857

Figure 1: Example comparisons showing NCD raw score using Borland's Pascal ZLib library (zlib 1.2.3). (Shakespearean lines taken from Quarto 1 and Folio 1, *Internet Shakespeare Editions* transcriptions.)

Scores below 0.5 appear to be strongly indicative of similarity, while those over 0.6 usually signify disparity; it is actually quite difficult to generate any score above 0.84, as shown by the final example, in which there are no points of similarity at all.

The prototype application takes two XML files as input, and performs the following steps:

- Identifies the target elements to be compared (this is currently hard-coded, but would ideally be based on user-specified XPath).
- Adds @id attributes to any of those elements which don't yet have them.
- Extracts the text of all the target elements, and normalizes it in a variety of ways (specified by the user).

- Compares each single text item with each other text item, and generates a comparison score for it.
- (Optional) Runs an additional contextualizing algorithm which modifies the original scores based on the scores of surrounding elements in the document (see below).
- Sorts the matches in ascending order of similarity score (best matches first).
- Presents each of these matches to the user for categorization as one of:
 - Corresponding (equivalent) items
 - Not corresponding, but an interesting relationship
 - No relationship at all
- Saves an XML file containing the scores for all matches, along with any categorization values chosen by the user.
- Saves copies of the input files with the added @id attributes, and also a CSV version of the score data for use in a spreadsheet program.

This correspondence data can be used to provide a component of a critical apparatus attached to a witness, linking it to corresponding lines in other witnesses.

Step 5 is an attempt to detect correspondences between lines in situations where a line has changed significantly, or perhaps even been replaced completely, but the lines around it still match closely. Normally, where lines differ, a high score will be generated; if the user chooses only to examine the lower scores in the search for correspondences, the link between these two lines may not be detected. The contextualizing algorithm massages the score of each match such that it is affected by the score of the lines around it; if the preceding and following lines have low scores, the score of the line is lowered so that it too may be detected more easily as "corresponding", even though it has been substantially changed. The contextualizing algorithm can be run many times if required, massaging the scores each time. We are still investigating the outcomes and value of this process.

These screenshots (Figures 2 and 3) show the prototype in use.



Figure 2: The main window of the prototype, showing the two input texts, and the Pre-comparison processing settings.

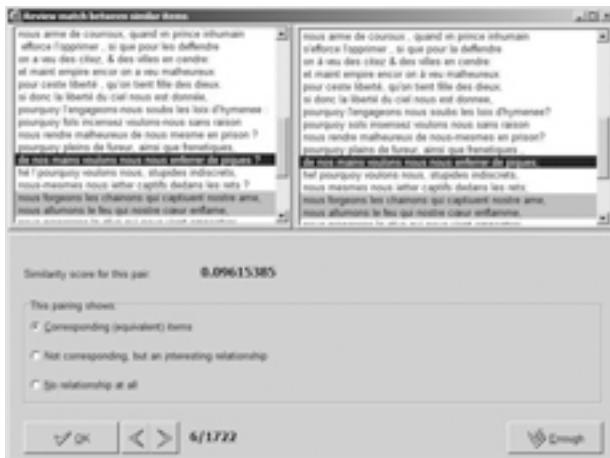


Figure 3: Reviewing matches between lines based on score.

This application is in many ways similar to the TEI-Comparator project created as part of the Holinshed Project. This is described in some detail in Cummings (2009). When writing our *Mariage* prototype at the beginning of 2009, I was unaware of the TEI-Comparator project; in addressing similar problems, we have arrived at remarkably similar solutions, especially in terms of process. The comparison algorithm used by TEI-Comparator, which is called Shingle Cloud, was developed by Arno Mittelbach, and uses a completely different process of comparison based on n-grams. Documentation for TEI-Comparator will be available soon; when it appears, I am looking forward to running tests to compare results between Shingle Cloud and the Universal Similarity Metric, and I will report the results in this paper.

The prototype application will probably now be ported to C++, to create a cross-platform application, and I also intend to create a standalone Java library that can be called from the command line

or from another Java application; perhaps it might be integrated into TEI-Comparator as an alternative comparison metric.

References

- Coste, Joël** (2008). 'Un regard médical sur la société française à l'époque d'Henri IV et de Marie de Médicis'. *XVIIe siècle*. **239**: 339-61.
- Cummings, James** (4 September 2009). "TEI-Comparator." Blog posting on *In my <element/>*. <http://blogs.oucs.ox.ac.uk/jamesc/2009/09/04/tei-comparator/>.
- Krasnogor, N. and D. A. Pelta** (2004). 'Measuring the similarity of protein structures by means of the universal similarity metric'. *Bioinformatics*. **20(7)**: 1015-1021. <http://bioinformatics.oxfordjournals.org/cgi/reprint/20/7/1015.pdf>.
- Li, Ming, Xin Chen, Xin Li, Bin Ma, and Paul Vitanyi** (2004). 'The Similarity Metric'. *IEEE Transactions on Information Theory*. **50(12)**: 3250-3264.
- Vitanyi, Paul** (2005). 'Universal Similarity'. Proc. ITW2005 - IEEE ITSOC Information Theory Workshop 2005 on Coding and Complexity. Rotorua, New Zealand, 29th Aug. - 1st Sept., 2005. <http://www.cwi.nl/~paulv/papers/itw05.pdf>.

Teasing Out Authorship and Style with T-tests and Zeta

Hoover, David L.

david.hoover@nyu.edu

New York University, USA

Most computational stylistics methods were developed for authorship attribution, but many have also been applied to the study of style. Investigating Wilkie Collin's *Blind Love* (1890), left unfinished at his death and completed by Walter Besant from a long synopsis and notes provided by Collins, requires both authorship attribution and stylistics. External evidence indicates that Besant took over after chapter 48 (Collins 2003), which provides an opportunity to test whether Besant was successful in matching Collins's style and to investigate the styles of Collins and Besant. This divided novel also facilitates the comparison of two computational methods: the T-test and Burrows's Zeta.

The t-test is a well-studied method for determining the probability of a difference between two groups arising by chance (a classic use in authorship and stylistics is Burrows 1992.) Here I use t-tests to identify words used very differently by Collins and Besant. After showing that those word frequencies accurately identify the change of authorship, I examine the words themselves for stylistically interesting characteristics.

I created a combined word frequency list for four novels by Besant and three by Collins, then deleted words occurring only once or twice, personal pronouns (too closely related to the number and gender of characters), all words with more than 90% of their occurrences in one text (almost exclusively proper names), and words limited to one author (required for t-testing). I divided the novels into 167 4,000-word sections, and performed t-tests for the remaining 6,600 words (using a Minitab macro). I cleaned up the results and sorted them on the p value in Excel (with another macro), and retained only the 1719 words with $p < .05$, about 1,000 for Collins and 700 for Besant (see <http://https://files.nyu.edu/dh3/public/ClusterAnalysis-PCA-T-testingInMinitab.html> for detailed instructions and the macros).

I tested these words on six new texts for each author, a novel and five stories for Besant and six novels for Collins. Beginning with the 500 most distinctive words for each author, I deleted a few words that were absent from these texts and used the remaining

993 words to perform a cluster analysis (Fig. 1). (To keep the graph readable, I divided the novels into 10,000- word sections, retaining only half the sections.) Obviously, these marker words are quite characteristic of the authors.

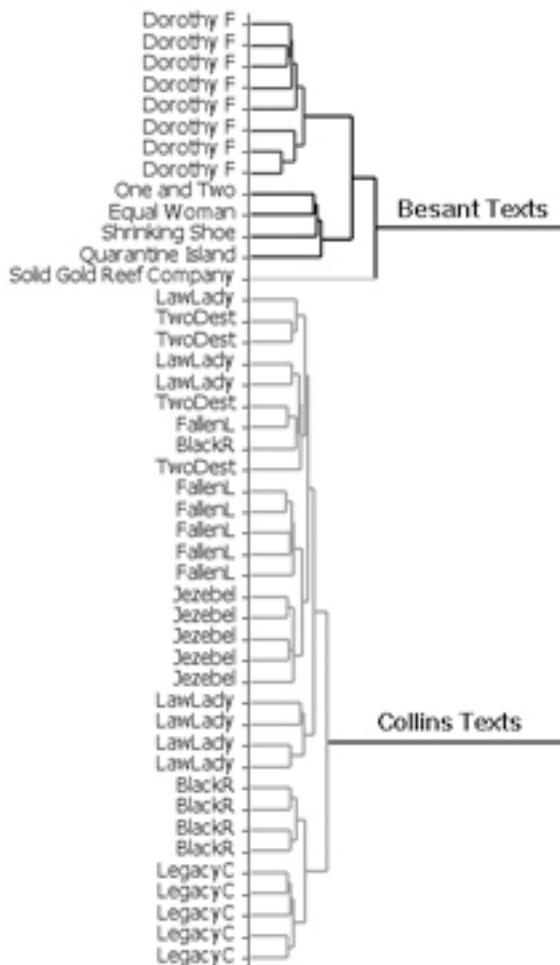


Fig. 1. Besant versus Collins: Cluster Analysis

When sections of *Blind Love* are tested along with the texts above, the authorship change after chapter forty-eight is starkly apparent (Fig. 2). This graph is based on the sums of the frequencies of the 500 most distinctive words for each author in each section. (The texts are divided into 1,000-word sections; only a few sections of the novels are shown; the frequencies of Collins's marker words are multiplied by -1 for clarity.) Although Besant was working from extensive notes, his style is distinctly different. Had we not known which was Besant's first chapter, these t-tested marker words would have easily located it.

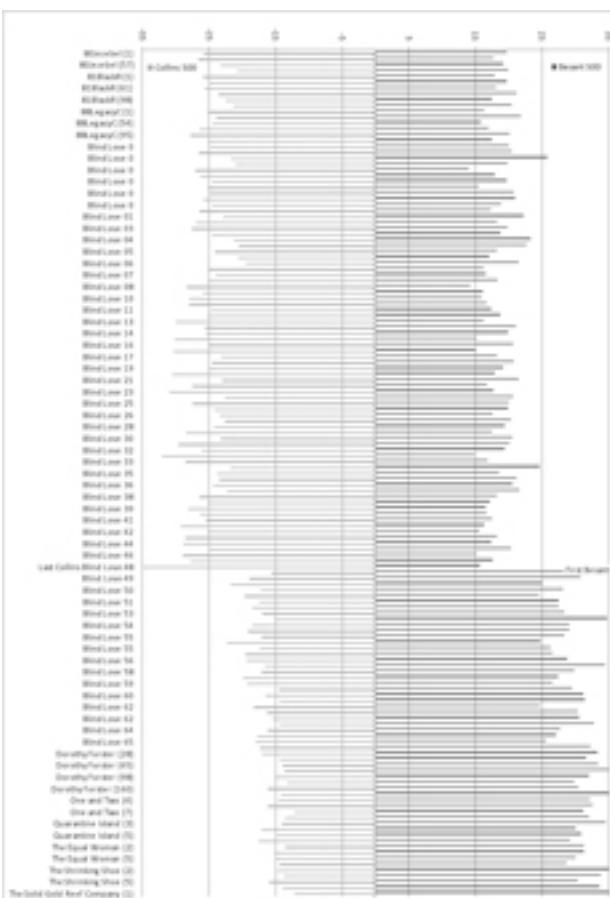


Fig. 2. Besant, Collins, *Blind Love*: T-tested Marker Words

Because the styles of Collins and Besant are so distinct, these marker words should also characterize them. Consider the twenty most distinctive words for each author:

Besant: *upon, all, but, then, and, not, or, very, so, because, great, thing, things, much, every, there, man, everything, is, well*

Collins: *answered, to, had, Mrs, on, asked, in, Miss, mind, suggested, person, resumed, excuse, left, at, reminded, creature, inquired, reply, when*

Obviously, more of Besant's words are high frequency function words, and many Collins words are related to speech presentation (*answered, asked, inquired, resumed, suggested, reply, and reminded*). The presence of *added, begged, declared, exclaimed, explained, expressed, muttered, rejoined, and said* as likely speech markers among the other Collins marker words, but only *gasped, groaned, murmured, replied, and stammered* for Besant, suggests they have different ways of presenting speech.

Sorting all of each author's marker words alphabetically immediately reveals word families that each author favors, as *thing, things, and everything* among the twenty most distinctive Besant words already suggests (*anything and nothing* are also Besant markers). His *every* and *everything* are

joined by *everybody* and *everywhere; anything* by *any* and *anywhere; nothing* and *not* by *never, no, nobody, none, and nor*; and *much* by *more, moreover, most, and mostly* among his markers. Collins's *answered* is joined by *answer, answering, and unanswerable*; and five of his twenty words are joined by two others: *ask, asked, asks; inquired, inquiries, inquiry; leave, leaving, left; person, personally, persons; suggest, suggested, suggestion*.

About 600 of the 1,700 distinctive words form groups favored by one author, but only about 175 form split groups, many of which fall into intriguing patterns. Collins uses more contractions, so *didn't, doesn't, and don't* are Collins words, but *did and does* are Besant words, and similarly for *must, need, should, and would* and their negative contractions. The singular and possessive forms of *brother, friend, sister, and son* are Collins's words and the plural forms are Besant's; the singular vs. plural pattern continues almost without exception in split noun groups. Verbs in *-ing* are Collins words and 3rd singular present forms Besant's. Finally, all nineteen cardinal number marker words are Besant's, including the numbers *one to ten* (note that Besant's preferred plural nouns often follow numbers). This extraordinary patterning may not seem particularly surprising, but, so far as I know, it has never been noticed before, and cries out for investigation.

Two problems with t-testing are its privileging of relatively uninteresting high-frequency words and its inability to cope with words absent from one author. John Burrows's Zeta addresses both of these problems (Burrows 2006). (The specific form used here was developed by Hugh Craig (Craig and Kinney, 2009); for an automated spreadsheet and instructions for performing Zeta analysis see <http://https://files.nyu.edu/dh3/public/TheZeta&IotaSpreadsheet.html>).

Zeta's simple calculation begins with the same novels and the same word frequency list used for the t-test, except that personal pronouns and words present in only one author are now included. Zeta is simply the sum of the proportions of Collins sections in which each word occurs and Besant sections in which it does not. Here *answered*, the most distinctive Collins word (as in the t-tests), has a Zeta score of 1.8, and is present in 89 of 90 Collins sections and absent from 65 of 77 Besant sections. The most distinctive Besant word is again *upon*, present in all 77 Besant sections and absent from 25 of 90 Collins sections, with a Zeta of 0.28. Below are the twenty most distinctive Zeta words (those also identified by t-testing in bold):

Besant: **upon, fact, presently, therefore, however, everything, real, whole, cannot, though, rich,**

none, thousand, except, fifty, ago, because, papers, also, twenty

Collins: **answered, Mrs, Miss, excuse, suggested, resumed, reminded, doctor, inquired, creature, notice, circumstances, tone, idea, temper, object, sense, feeling, governess, impression**

As noted above, Zeta marker words are less frequent than t-tested words. Only two Zeta marker words rank in the top 100 in the novels, compared to 20 of the t-tested words. About 3/4 of the 1000 t-tested marker words are also among the 1000 Zeta markers. Among the 2000 Zeta words are 275 words occurring in only one author; 59 form new single-author families, 27 join existing single-author families, and only 21 form split families.

The Zeta words also effectively detect the change of authorship in *Blind Love*. In the scatter graph in Fig. 3, the axes show the percentages of all the word types (unique words) in each section that are Besant or Collins marker words (longer texts are divided into 4000-words sections; the labels for even-numbered Collins sections of *Blind Love* are removed; only a few sections of other novels are included). Note how distinct Besant's chapters of *Blind Love* are from Collins's, though many of them are pulled toward Collins. This graph also includes *The Case of Mr. Lucraft* (Case in bold), jointly written by Besant and James Rice; it suggests, as has been argued (Boege 1956: 251-65), that Besant did most of the actual writing.

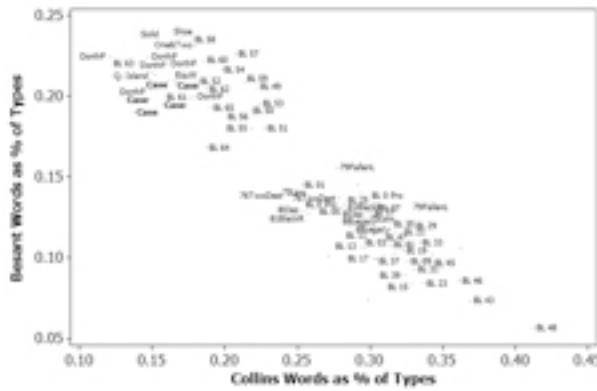


Fig. 3. Besant, Collins, *Blind Love*: Zeta Analysis

T-tests and Zeta analysis are both effective authorship attribution methods that produce lists of characteristic vocabulary for the authors being compared. Both identify morphological and semantic families of words and uncover extraordinarily consistent patterns and puzzling inconsistencies that suggest new directions for literary and stylistic analysis.

References

- Boege, F. (1956). 'Sir Walter Besant: Novelist. Part One'. *Nineteenth-Century Fiction*. **10**: 249-280.
- Burrows, J. F. (1992). 'Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information'. *LLC*. **7**: 91-109.
- Burrows, J. F. (2006). 'All the Way Through: Testing for Authorship in Different Frequency Strata'. *LLC*. **22**: 27-47.
- Collins, W. (2003). *Blind Love*. Bachman, M., Cox, D. (eds.). Peterborough, Ont.: Broadview Press.
- Collins, W. (2009). *Blind Love*. London: Chatto & Windus. <http://ia311528.us.archive.org/0/items/blindlove00colluoft/blindlove00colluoft.pdf> (accessed 18th March 2009).
- Craig, H., Kinney, A., (eds.) (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

A New Digital Method for a New Literary Problem: A Proposed Methodology for Briding the "Generalist" - "Specialist" Divide in the Study of World Literature

Howell, Sonia

sonia.howell@nuim.ie

School of English and Media Studies, National University of Ireland, Maynooth, Maynooth Co. Kildare, Ireland

Keating, John G.

john.keating@nuim.ie

An Foras Feasa: The Institute for Research in Irish Historical and Cultural Traditions, National University of Ireland, Maynooth, Maynooth Co. Kildare, Ireland

Kelleher, Margaret

margaret.kelleher@nuim.ie

An Foras Feasa: The Institute for Research in Irish Historical and Cultural Traditions, National University of Ireland, Maynooth, Maynooth Co. Kildare, Ireland

This paper is situated within the debate between "specialist" and "generalist" methods of analysis in the study of world literature. It is argued that a systemic linguistic discourse analysis of appropriately encoded text passages can provide a methodology which can be utilised to interrogate the national and international demarcations of comparative literary analysis. A case study consisting of a textual analysis of the dialogical relationship between patient and therapist in a "factional", i.e. works of fiction, which draw upon historical fact. Irish and English novel is provided. The benefits of the results yielded to current understandings of national literature and definitions of world literature are discussed.

In 1827, the German poet, Johann Wolfgang von Goethe, declared to his young disciple Johann Peter Eckermann that "National literature is now a rather unmeaning term; the epoch of world literature is at hand, and everyone must strive to hasten its approach" (Eckermann, cited in [Damrosch, 2003b, p. 1]). History informs us however that Goethe was premature in his heralding of a new age of "postnational" literature, as up until recently all literatures tended to have been studied along national lines [Dimock, 2006, pp. 2-5]. Yet in more

recent decades, nations and, by extension, "national" literatures have become increasingly under threat in their sovereignty over all elements of human life due to the homogenising and heterogenising effects of globalisation. Globalisation is defined by Malcolm Waters as being "a social process in which the constraints of geography on social and cultural arrangements recede and in which people become increasingly aware that they are receding" [Waters, 1995, p. 3]. It is not surprising therefore, that in an age where national boundaries, both physical and mental, are become increasingly insignificant and blurred, we find a renewed interest in Goethe's concept of *Weltliteratur*.

As with all things new, the emerging discipline of world literature has evoked fear and reservations among literary scholars. According to David Damrosch, the possibility of recognizing the ongoing, vital presence of the national within the life of world literature poses enormous problems for the study of world literature [Damrosch, 2003a, p. 514]. Thus the field tends to be divided into "specialists", those who are concerned with national literatures, and "generalists", those who are interested in studying global patterns. But instead of this either/or method, Damrosch maintains that what is need is a method that can mediate between broad and often reductive overviews and intensive, but often atomistic close readings [p. 519]. As Franco Moretti argues, "we must find a way to combine the individual who reads a single work with great collective efforts and vision" [Sutherland, 2006, Monday 9 January, 2006]. This paper argues that a combination of Systemic Functional Linguistics and Digital Humanities offers one way whereby this may be achieved.

The complexities involved in the interpreting of "literary language" for electronic media have perhaps posited the greatest deterrent for literary scholars in embracing digital humanities to date. The pioneering work of scholars such as Willard McCarty and Jerome McGann (among others) unfortunately remain the exception among their peers in the field of literary studies. The majority within the discipline retain the fear that computer based analysis of texts can only reveal "broad sweeping patterns" within literary works. Interestingly, this fear echoes the reservations that are held about the theoretical methods deployed by "generalists" in the study of world literature. Contrary to both these fears, this paper will argue that computer-based literary research can be utilized to provide both a means of analysis for the specificities of national literatures, while also serving as a tool for carrying out comparative textual analysis at an international scale. We wish to present to the following methodology to the community.

1. Case Study: The Patient-Therapist Relationship in Literature

This project will consist of an analysis of a passage of dialogue between patient and therapist in Sebastian Barry's *The Secret Scripture* (2008) and Pat Barker's *Regeneration* (1995). For the purpose of this study, we have chosen a novel by an Irish writer and an English writer respectively. Given that this methodology is in its infancy, it is presumed best to begin with two texts written in the same language and which originate from countries of similar cultural systems. *The Secret Scripture* is set in present day Ireland but the narrative is made up of a double narrative; the personal recollections of Roseanne Clear, who was incarcerated in a mental institution during the mid twentieth century, and the account by the psychiatrist, Dr. Greene, of his own investigation into Roseanne's admittance into the hospital. *Regeneration* is based on the real-life experiences of British army officers being treated for shell shock during World War I at Craiglockhart War Hospital in Edinburgh. Its narrative relays the treatment of soldiers suffering mental break down. It is shaped predominately around the discussions which the psychiatrist, Dr. Rivers, has with a number of patients within the asylum in which he works. Both novels are centered around events which have caused psychological distress to the individual characters, but which have also caused what is known as 'cultural trauma'¹ to the nations in which they are set.

2. Methodology

Our method of analysis is based on Systemic functional linguistics (SFL) which is a model of grammar that was developed by Michael Halliday in the 1960s [Halliday, 1976, Halliday, 2004]. It is part of a broad social semiotic approach to language called systemic linguistics. The term *systemic* refers to the view of language as "a network of systems, or interrelated sets of options for making meaning". The term *functional* indicates that the approach is concerned with meaning, as opposed to formal grammar, which focuses on word classes such as nouns and verbs, typically without reference beyond the individual clause. Systemic-Functional Linguistics (SFL) is a theory of language centred around the notion of language function. SFL places the function of language as central (what language does, and how it does it), in preference to more structural approaches, which place the elements of language and their combinations as central. Specifically, it begins with a social context, and looks at how language both acts upon, and is constrained by, this social context. In the model,

and methodology, particular aspects of a given social context (such as the topics discussed, the language users and the medium of communication) define the meanings likely to be expressed and the language likely to be used to express those meanings. Since language is viewed as semiotic potential, the description of language is a description of choice. Systemic linguists examine the choices language users can make in a given setting to realise a particular linguistic product (the available choices depend on aspects of the context in which the language is being used). By examining the different choices for the discourse between characters in similar social contexts in different texts, we believe that it may be possible to identify, or describe, the features associated with national literatures that address the problem described here.

Our approach is essentially a discourse analysis rather than textual analysis, although it relies on encoded text for comparative analysis. Specifically, we draw on existing SFL models of sociolinguistic and cognitive approaches to doctor-patient discourse [Todd and Fisher, 1993, Togher, 2001]. These models (i) analyse the patterns of talk that are produced by the situational demands of the particular setting, (ii) provide a detailed examination of the interplay of language use in this organisational context of health care delivery, and (iii) examine the production of doctor-patient communication, and (iv) examine the relationship between social structure and social interaction, and explore the relationship between power and resistance. The discourse analyses will utilise custom developed software that analyses encoded passages of text from the novels. A number of encoding schemes exist and provide mechanisms for encoding linguistic data, for example, the Text Encoding Initiative (TEI) [Sperberg-McQueen and Burnard, 1994] or the more recent XCES, based on the Corpus Encoding Standard (CES) which is a part of the EAGLES Guidelines developed by the Expert Advisory Group on Language Engineering Standards (EAGLES). EAGLES provides a set of encoding standards for corpus-based work in natural language processing applications [Ide et al., 2000].

3. Discussion

Through the investigation into past experiences in an attempt to divulge facts about the present condition, the patient-therapist relationship has enormous significance in "factional" novels, given their link to cultural memory and cultural trauma. For example, the revelation of the occurrence of nominalisation in the dialogue between patient and therapist provides textual evidence as to whether the patient successfully 'moves on' from the traumatic past experience. Nominalisation is an example of *grammatical metaphor*, the term given when one

grammatical class is substituted for another, for example, replacing *he departed* with *his departure*. The lexical items are the same, but their place in the grammar has changed. In this example, the meaning expressed by an individual differs in that one form identifies a transient event while the other is more permanent. Attention is also paid to the gender of the characters. The case-study will provide (i) an example of a recurring literary structure of focused dialogue, (ii) one which has a general referent in therapeutic terms: literature of trauma etc., and (iii) but also a dialogue which gathers particular meaning from its specific national and cultural context of conflict (in the case of our chosen novels, an emerging independent Ireland and immediate post World War I Britain).

We demonstrate how the findings of our study can be utilized to provide a specialist analysis of the texts within their respective national literatures of Ireland and Britain by applying them to answer a specific literary question about the texts in their national contexts. Having done so, we then illustrate that our findings also elucidate what can be achieved by generalist studies in the context of world literature by comparing the results to provide a commentary on variations in the artistic treatment of cultural trauma.

4. Conclusion

Our case study tests a methodology, which is concerned with analyzing a ‘common-denominator’ (the patient-therapist relationship) between texts. This opens up the two novels under examination to a useful comparative reading as works of world literature, while also yielding useful results to study of the texts within their respective national literatures. In so doing it (i) introduces a new methodology to the academic community and (ii) demonstrates the application of the results produced by said methodology to the answering of a specific literary question. The ultimate aim of the model is to provide a digital linguistic tool that supports individual and collaborative projects in the study of world literature, thus assuaging the needs of both ‘specialists’ and ‘generalists’. In so doing, it offers a solution to one of the problems present in the study of world literature while simultaneously advancing the use of digital humanities in literary studies.

References

- Alexander, J. C., Eyerman, R., Giesen, B., Smelser, N. J., Sztompka, P.** (2004). *Cultural Trauma and Collective Identity*. London: California University Press.
- Damrosch, D.** (2003a). *What is World Literature?*. Princeton and Oxford: Oxford University Press.
- Damrosch, D.** (2003b). 'World literature, national contexts'. *Modern Philology*. **100**: 512–531.
- Dimock, W. C.** (2006). *Through Other Continents: American Literature Across Deep-Time*. Princeton, NJ: Princeton University Press.
- Halliday, M. A. K.** (1976). *Cohesion in English*. London: Longman.
- Halliday, M. A. K.** (2004). *An introduction to functional grammar*. London: Arnold.
- Ide, N., Bonhomme, P., Romary, L.** (2000). 'Xces: An xml-based encoding standard for linguistic corpora'. *LREC-2000*. Athens, Greece, pp. 825–830.
- Sperberg-McQueen, C. M., Burnard, L.** (1994). *Guidelines for the encoding and interchange of machine-readable texts*. Chicago and Oxford: Text Encoding Initiative.
- Sutherland, J.** (Monday, 9 January, 2006). 'The ideas interview: Franco moretti'. *The Guardian*.
- Todd, A. D., Fisher, S.** (1993). *The Social Organization of Doctor-Patient Communication*. London: Ablex Publishing.
- Togher, L.** (2001). 'Discourse sampling in the 21st century'. *Journal of Communication Disorders*. **34**: 131–150.
- Waters, M.** (1995). *Globalization*. London: Routledge.

Notes

1. Neil J. Smelser defines cultural trauma as being a memory accepted and publicly given credence by a relevant membership group and evoking an event or situation which is a) laden with negative effect, b) represented as indelible, and c) regarded as threatening a society's existence or violating one or more of its fundamental cultural presuppositions (Smelser cited in [Alexander et al., 2004, p. 44])

"Litmap": Networked Narratives

Hui, Barbara

barbara.hui@gmail.com

UCLA, USA

This paper examines the spatiality of three contemporary literary narratives using a digital humanities approach. By this I mean a few things: firstly, I regard spatiality as a complex and dynamic historical dimension on par with temporality, and not just as a static, passive container in which events independently transpire. Secondly, I am interested in examining not only space and place as represented in texts, but also the spatiality of the texts themselves, i.e., the materiality of language. Thirdly, I have built the *Litmap* digital mapping platform (<http://barbarahui.net/litmap>) for the purpose of visualizing space and place in/of texts, which I use in conjunction with traditional close reading methods in order to carry out my scholarship.

The definition of spatiality I employ follows from arguments made by spatial theorists including Henri Lefebvre, David Harvey, Doreen Massey, Edward Said and Edward Soja, who push for an understanding of space and place as socio-historically produced rather than somehow existing *a priori*. I argue further that networked spatiality is a prevalent trope, organizing principle, and way of understanding the world in contemporary texts. I show how this presents itself in the narratives I examine (in quite a different way in all three) and is a particularly useful, even crucial inroad into understanding them. My assertion is that the three texts at hand can be characterized as displaying three kinds of topographical networks:

- In W.G. Sebald's *Rings of Saturn* (1997), geographical places are connected to each other via a historical network of events, and the nodes of the network are primarily man-made architectural structures.
- In Emine Sevgi Özdamar's *Seltsame Sterne starren zur Erde* (2008), geographical places are connected to one another via the transnational migrations of people, and the nodes of the network are these moving embodied subjects themselves.
- In Steven Hall's *Raw Shark Texts* (2007), language, thought, and memory are material and have spatial dimensions. Places are connected to each other via these material traces, and the nodes of the network, which are constituted by human subjects and their linguistic traces, are

ephemeral and unstable, with "un-space" figuring as an otherworldly yet very real dimension in the narrative's spatial imaginary. In addition, the text of *Raw Shark Texts* itself is figured as a material body of language and textual image, with patterned connections running throughout the book.

Both the core observations listed above and the sub-arguments presented in the thesis were arrived at via a combination of *Litmap*-based and traditional print-based research methodologies. The current *Litmap* interface displays a map image of the Earth, with place names and corresponding information from each text keyed to that location's coordinates on the map. In the case of *Rings of Saturn*, this allows for a fairly complex mapping since the nodes of the network in that narrative correspond to unambiguous geographical place names and locations. In the case of Özdamar's and Hall's texts, however, this becomes increasingly challenging as space and place become more subjective and fluid, requiring new and creative ways of visualizing data. The use of digital media to map literature is thus useful both for revealing what it can and can't do, and I argue it is important to recognize both the strengths and constraints of the medium as we continue exploring this new area of research.

Moving forward, I plan to develop and extend the *Litmap* platform both in order to better address the crucial issues of how to visualize ambiguous data, and also to improve upon the existing functionality for searching, filtering, and browsing. The database underlying the current system is flexible and extensible enough to accommodate information from far more narratives, and I intend to enable users to upload other books for teaching and research. Once a large corpora of texts is uploaded, this will open up the ability to search across time and space and do other macro analyses of literature. Perhaps most of all, however, I look forward to making *Litmap* a truly collaborative project. My hope is to assemble a team of colleagues who will be invested in working together on creative technical and design solutions for the platform.

References

- Hall, Steven** (2007). *The Raw Shark Texts*. New York: Canongate.
- Özdamar, Emine Sevgi** (2008). *Seltsame Sterne starren zur Erde: Wedding - Pankow 1976/77 2003*. Köln: Kiepenhauer & Witsch.
- Sebald, W.G.** (1998). *The Rings of Saturn: An English Pilgrimage*. Hulse, Michael (ed.). New York: New Directions Books.

The Open Annotation Collaboration: A Data Model to Support Sharing and Interoperability of Scholarly Annotations

Hunter, Jane

j.hunter@uq.edu.au

University of Queensland

Cole, Tim

t-cole3@illinois.edu

University of Illinois, Urbana-Champaign

Sanderson, Robert

azaroth42@gmail.com

Los Alamos National Laboratory

Van de Sompel, Herbert

hvdsomp@gmail.com

Los Alamos National Laboratory

This paper presents the outcomes to date of the annotation interoperability component of the Open Annotation Collaboration (OAC) Project.¹ The OAC project is a collaboration between the University of Illinois, the University of Queensland, Los Alamos National Laboratory Research Library, the George Mason University and the University of Maryland. OAC has received funding from the Andrew W. Mellon Foundation to develop a data model and framework to enable the sharing and interoperability of scholarly annotations across annotation clients, collections, media types, applications and architectures. The OAC approach is based on the assumption that clients publish annotations on the Web and that the target, content and the annotation itself are all URI-addressable Web resources. By basing the OAC model on Semantic Web and Linked Data practices, we hope to provide the optimum approach for the publishing, sharing and interoperability of annotations and annotation applications. In this paper, we describe the principles and components of the OAC data model, together with a number of scholarly use cases that demonstrate and evaluate the capabilities of the model in different scenarios.

1. Introduction and Objectives

Annotating is both a core and pervasive practice for humanities scholarship. It is used to organize, create and share knowledge. Individual scholars use it when reading, as an aid to memory, to add commentary,

and to classify documents. It can facilitate shared editing, scholarly collaboration, and pedagogy. Although there exists a plethora of annotation clients for humanities scholars to use (Hunter 2009) - many of these tools are designed for specific collection types, user requirements, disciplinary application or individual, desktop use. Scholars are also confronted with having to learn different annotation clients for different content repositories, have no easy way to integrate annotations made on different systems or created by colleagues using other tools, and are often limited to simplistic and constrained models of annotations. For example, many existing tools only support the simplistic model in which the annotation content comprises a brief unformatted piece of text. Many tools conflate the storage of the annotations and the target being annotated.

Frameworks for annotation reference are inconsistent, not coordinated, and frequently idiosyncratic, and the constituent elements of annotations are not exposed to the Web as discrete addressable resources, making annotations difficult to discover and re-use. The lack of robust interoperable tools for annotating across heterogeneous repositories of digital content and difficulties sharing or migrating annotation records between users and clients – are hindering the exploitation of digital resources by humanities scholars. Hence the goals of the Open Annotations Collaboration (OAC) are:

- To facilitate the emergence of a Web and Resource-centric interoperable annotation environment that allows leveraging annotations across the boundaries of annotation clients, annotation servers, and content collections. To this end, annotation interoperability specification consisting of an Annotation Data Model will be developed.
- To demonstrate through implementations an interoperable annotation environment enabled by the interoperability specifications in settings characterized by a variety of annotation client/server environments, content collections, and scholarly use cases.
- To seed widespread adoption by deploying robust, production-quality applications conformant with the interoperable annotation environment in ubiquitous and specialized services and tools used by scholars (e.g., JSTOR, Zotero, and MONK).

In the remainder of this paper we describe related efforts that have informed the development of our Annotation Data Model. We then describe the data model itself that lays a foundation for follow-on work involving demonstrations and reference implementations that exploit real-world repositories such as *JSTOR*, *Flickr Commons*, and *MONK* and

leverage existing scholarly annotation applications such Zotero, Pliny and Co-Annotea.

2. Related Work

Despite the vast body of work regarding annotation practice, annotation models, and annotation systems, little attention has been paid to interoperable annotation environments. The few efforts in this realm to date comprise:

- RDF-based Annotea developed by Kahan and Koivunen (Kahan et al., 2001);
- Agosti's "A Formal Model of Annotations of Digital Content" (Agosti and Ferro, 2007);
- SANE Scholarly Annotation Exchange;
- OATS (The Open Annotation and Tagging System (Bateman et al., "OATS: The Open Annotation and Tagging System").)

An analysis of these existing models reveals that on the whole, they have not been designed as Web-centric and resource-centric, or that they have modeling shortcomings that prevent any existing resource from being the content or target of an annotation and from giving an annotation independent status as a resource itself. Further requirements that we have identified that these approaches fail to fully support include:

- Resources of any media type can be Annotation Content or Targets;
- Annotation Targets or Content are frequently segments of Web resources;
- The Content of a single annotation may apply to multiple Targets or multiple annotation Contents may apply to a single Target;
- Annotations can themselves be the Target of further Annotations.

3. The OAC Data Model

By exploiting the Web- and Resource-centric approach to modelling annotations, we leverage existing standards and facilitate the interoperability of annotation applications. In the OAC model, an Annotation is an Event initiated at a date/time by an author (human or software agent). Other entities involved in the event are the Content of the Annotation (aka Source) and the Target of the Annotation. The model assumes that the core entities (Annotation, Content and Target) are independent Web resources that are URI-addressable. This approach simplifies and decouples implementation from the repository. An essential aspect of an annotation is the (implicit or explicit) expression of "annotates" relationship between the Content and

the Target. The model allows for Content and Target of any media type and the Annotation, Content, and Target can all have different authorship. In situations where the annotation Content or Target is a segment or fragment of a resource (e.g., region of an image), we will draw on the work of the W3C Media Fragments Working Group to specify the fragment address. Figure 1 illustrates the alpha version of the OAC data model.

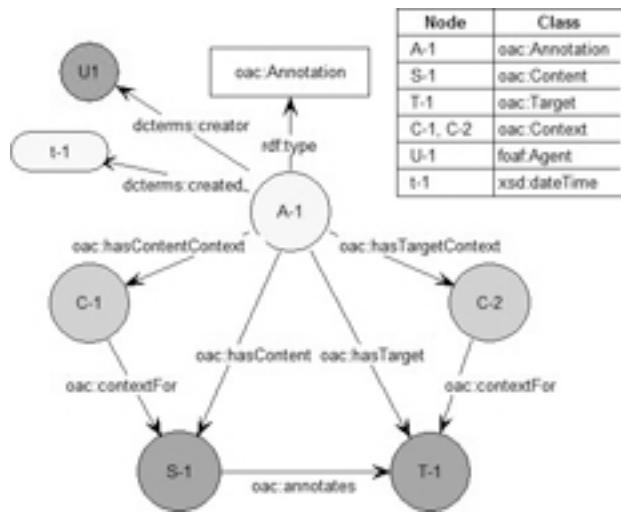


Fig. 1. The Alpha OAC Data Model

4. Use Cases

In order to evaluate and demonstrate the feasibility of the OAC Data Model, an initial set of use cases has been developed that are representative of a range of common scholarly practices involving annotation. This preliminary set is available from the OAC Wiki as OAC User Narratives/Use Cases² and includes:

- Citation of Non Printed Media
- Commentary on Remote Resources
- Shared Annotations Across Interfaces
- Harvesting, Aggregating, Ranking and Presenting Annotations from Multiple Sites
- Annotating Relationships Between Multiple Mixed-Media Resources
- Annotations which Capture Net chaining Practices
- Annotations with Compound Targets

For example, Figure 2 illustrates a scholarly annotation example involving multiple targets, in which a scholar is making a comment on the differences between segments in scholarly editions of the poem "The Creek of the Four Graves" by Charles Harpur.



Fig. 2. Annotating the differences between two scholarly editions in AusLit

Figure 3 below illustrates the corresponding OAC model for the use case in Figure 2 in which a single annotation Content applies to two Target resources.

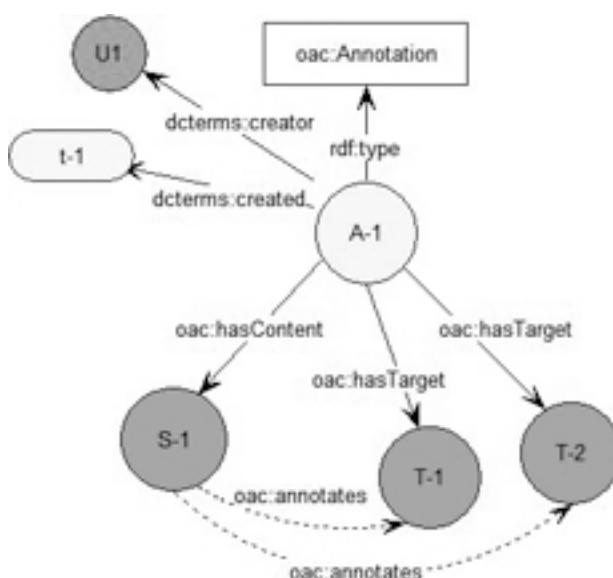


Fig. 3. OAC Model for the example in Figure 2

5. Discussion and Conclusions

The proposed OAC Data Model will enable the sharing and discovery of annotations beyond the boundaries of individual solutions or content collections, and hence will allow for the emergence of value-added cross-environment annotation services. It will also facilitate the implementation of advanced end-user annotation services targeted at humanities scholars that are capable of operating across a broad range of both scholarly and general collections. Furthermore, it will enable customization of annotation services for specific scholarly communities, without reducing

interoperability. The proposed work will also enable more robust machine-to-machine interactions and automated analysis, aggregation and reasoning over distributed annotations and annotated resources. By grounding our work in a thorough understanding of Web-centric interoperability and embedded models implemented by existing digital annotation tools and services, we create an interoperable annotation environment that will allow scholars and tool-builders to leverage prior tool development work and traditional models of scholarly annotation, while simultaneously enabling the evolution of these models and tools to make the most of the potential offered by the Web environment.

Acknowledgments

The Open Annotations Collaboration (OAC) is funded by the Andrew W. Mellon Foundation. The authors would also like to acknowledge the valuable contributions to this work made by: Neil Fraistat, Doug Reside, Daniel Cohen, John Burns, Tom Habing, Clare Llewellyn, Carole Palmer, Allen Renear, Bernhard Haslhofer, Ray Larsen, Cliff Lynch and Michael Nelson. Figure 2 is courtesy of Anna Gerber, Senior Software Engineer on the Aus-e-Lit project.

References

- Hunter J.** (2009). 'Collaborative Semantic Tagging and Annotation Systems'. *Annual Review of Information Science and Technology (ARIST)*. ASIST.
- Kahan, J., Koivunen, M.** (2001). 'Annotea: An Open RDF Infrastructure for Shared Web Annotations'. *Procs of the 10th International conference on the World Wide Web*. Pp. 623-632.
- Agosti, M., Ferro, N.** (2007). 'A Formal Model of Annotations of Digital Content'. *ACM Transactions on Information Systems*. 1.
- SANE Scholarly Annotation Exchange*. <http://www.huygensinstituut.knaw.nl/projects/sane/>.
- Bateman, S., Farzan, R., Brusilovsky, P., McCalla, G..** *OATS: The Open Annotation and Tagging System*. <http://fox.usask.ca/files/oats-lornet.pdf>.
- Sanderson R., Van de Sompel H..** *Open Annotation Collaboration Alpha Data Model Summary*. http://www.openannotation.org/documents/OAC-Model_UseCases-alpha.pdf.

Notes

- [1. http://www.openannotation.org/](http://www.openannotation.org/)

2. <http://https://apps.lis.illinois.edu/wiki/display/openannotation/OAC+User+Narratives---Use+Cases>

A corpus approach to cultural keywords: a critical corpus-based analysis of ideology in the Blair years (1998-2007) through print news reporting

Lesley Jeffries

l.jeffries@hud.ac.uk

The University of Huddersfield, UK

Brian David Walker

b.d.walker1@lancaster.ac.uk

The University of Huddersfield, UK

This paper will report on a corpus-based study of the cultural keywords (in the Raymond Williams' sense) via the analysis of key-words (in the corpus/statistical sense) of newspaper reporting in the years since Labour came to power. The project demonstrates that certain lexemes (or lexical strings) gain currency in relatively short historical periods and may take on political importance.

This project assesses the ideological landscape during the years of the New Labour project by extracting the cultural keywords of the time, and demonstrating their evolving meanings in the commentary provided by the print media.

The project takes inspiration from Raymond Williams' book ([1975] 1983) *Keywords* which attempted to sum up the ideology of the post-war years. Williams chose a set of words which he thought had taken on particular meanings in that period, and wrote an informed but ultimately anecdotal commentary on each one. Like Williams, we begin with a hypothesis that some words (such as, for example, *radicalisation*, *choice*) have both increased in usage and polarised in their meaning since 1998.

Unlike Williams, this project pursues a rigorous approach to the discovery of *which* words characterise the period under investigation, using two corpora of newspaper data and computer tools. This enables us to make a comprehensive investigation and an objective assessment, including use and meaning, of the cultural keywords of the Blair years

The project is primarily corpus-based, but with a strong qualitative focus, using an approach to studying textually-constructed meanings of words and other linguistic items which recognises both their place in a relatively stable system of language,

and their capacity to take on additional meaning in specific contexts of time and place.

1. Background

Our project links the corpus linguistic notion of key-words to earlier work into the ‘emergent meaning’ of individual lexical items (see Jeffries 2003, 2006 and 2007).

Jeffries (2003) investigated the meaning of water, in the context of the Yorkshire water crisis of 1995. Jeffries (2006) investigated the speech act of apology, in particular news commentators’ view of Blair’s putative apology for the Iraq war. Jeffries (2007) was a much more extensive consideration of the way in which the female body was constructed by women’s magazines in 2004. This larger study developed a system of describing textual meaning which draws on Hallidayan approaches to systems of linguistic form and meaning applying his combined semantic and syntactic view of textual meaning to other functions such as the construction of opposites.

The current project was designed in the spirit of Critical Discourse Analysis, in particular the work of Fairclough whose work on ideology in language, and specifically the language of New Labour (Fairclough 2000) influences the approach taken here. However, the methods used in this project are closer to corpus stylistics in that they are text-analytic and at least in some of the stages, computer-assisted and corpus-driven. Work already carried out in this area (see for example McIntyre and Walker 2010) showed that corpus approaches and tools, in particular Wmatrix, can successfully be applied to textual analysis. Baker and McEnery (2005) and Baker and Gabrielatos (2008) are also influential on this project because this work has paralleled Jeffries’ work in looking at sets of texts from a particular time period to demonstrate political ideologies in news texts.

The project also reflects renewed interest in cultural keywords in the Williams sense, with a recent special issue of *Critical Quarterly* (2007) devoted to the subject, and Durant’s (2006) related article which suggests that “[...] the development of electronic search capabilities applied to large corpora of language use [...] encourages renewed attention to cultural keywords.” (Durant 2006). This project effectively takes up that suggestion.

2. Research questions

- What are the key-words for the years 1998–2007, as evidenced in the British press and can they be identified as cultural keywords?

- Have they developed meanings specific to this period and have these meanings evolved within the period?

3. Methodology

The project focuses on news texts from 1998 to the end of 2007. A corpus of comparable data from three national daily newspapers (*The Guardian*, *The Independent*, and *The Times*) was assembled from a large, on-line newspaper database. This database represents a very rich and potentially overwhelming amount of data (100s of millions of words). However, our project had very limited timescales and we found it necessary to carefully control the amount of data that we collected. This was because: (i) downloading selected articles from the database is largely a manual and fairly time consuming process; and (ii) in its raw form each downloaded article contained structured extra-textual details (headers containing titles, dates and so forth), random intra-textual information (such as journalist’s email addresses) and corruptions. This extra text and corrupted data had to be removed from and amended in each of the downloaded files: a laborious process which consumed a lot of project time. We also found that the corpus tools we used for data manipulation struggled when presented with files of more than one million words. Consequently, we took a structured sampling approach, choosing a week from the politically ‘busy’ month of September (party conferences), and collected selected news-related items from these weeks. The resulting corpus was approximately 2.3 million words, which we anticipated would be sufficient to answer our research questions. A comparison corpus was built along similar lines using newspaper data from the five year period prior to 1997 (the Major years).

The corpus was automatically analysed, in the first instance, using Wmatrix (Rayson 2008), which is a relatively new corpus tool that can calculate keyness (using Log-likelihood) at the word level (key-words), at the grammatical level (key-POS), and the semantic level (key-concepts). The present study uses just the key-word output.

To address the qualitative aspect of the research questions, this investigation included the following considerations:

- Do the collocations of the key-word demonstrate particular nuances of meaning?
- How does the semantico-syntactic behaviour of the key-word demonstrate meaning specific to the context?
- Does the key-word enter into any unconventional lexical relations (e.g. of opposition)?

- Is the key-word associated with any modal or negated text worlds?

4. Results

The key-words, generated from the comparison of our corpora, that we consider to be the important cultural keywords from the Blair years are as follows:

No.	Keyword	Associated key-words
1	Terror	Terrorism, terrorist(s), attacks, atrocities, threat
2	Global	Globalisation, world, international
3	Spin	spun
4	Reform	progressive, radical, modernise(d) / er(s) / ation
5	Choice	
6	Respect	

Items in the 'keyword' column are the main items used in our investigation and the terms that we consider to be culturally significant. The items in the third column are key-words resulting from our corpus comparison which are related to individual (cultural) keywords and which, we hypothesise, form a network of meaning. These are still to be fully investigated and we do not report on them in this paper.

For each keyword we provide a more detailed quantitative commentary using concordance and collocation data. Our major findings, though rigorous and replicable, are qualitative, and provide the basis of both detailed linguistic commentaries on each key-word and could also provide the foundation for more general popular essays not dissimilar to those provided by Williams, but with more clarity about their provenance. There will not be time to discuss all our findings, but our paper will report on some of the quantitative data and focus qualitatively on 'spin'.

confirmed, asserted, qualified, changed". *Critical Quarterly*. **48:1**: 1–26.

Fairclough, N. (2000). *New Labour, New Language*. London: Routledge.

Jeffries, Lesley (2003). 'Not a drop to drink: Emerging meanings in local newspaper reporting of the 1995 water crisis in Yorkshire'. *Text - Interdisciplinary Journal for the Study of Discourse*. **23 (4)**: 513–538.

Jeffries, Lesley (2006). 'Journalistic Constructions of Blair's 'Apology' for the Intelligence Leading to the Iraq War'. *Language in the Media: Representations, Identities, Ideologies. Advances in Sociolinguistics*. London: Continuum.

Jeffries, Lesley (2007). *Textual Construction of the Female Body A Critical Discourse Approach*. Basingstoke.

McIntyre, D., Walker, B. (2010). "How can corpora be used to explore the language of poetry and drama?". *The Routledge Handbook of Corpus Linguistics*. McCarthy, M., O'Keefe, A. (eds.). Abingdon: Routledge.

Rayson, P. (2008). *Wmatrix: a web-based corpus processing environment*, Computing Department, Lancaster University <http://www.comp.lancs.ac.uk/ucrel/wmatrix/>.

Williams, R. (1983). *Keywords (2nd Ed.)*. London: Fontana.

References

Adamson, S., Durant, A (2007). *Critical Quarterly*. **49,1**.

Baker, P., McEnery, A. (2005). 'A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts'. *Language and Politics*. **4:2**: 197–226(30).

Baker, P., Gabrielatos, C. (2008). 'Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996–2005'. *Journal of English Linguistics*. Forthcoming.

Durant, A. (2006). 'Raymond Williams's Keywords: Investigating Meanings "offered, felt for, tested,

The Modern Art Iraq Archive (MAIA): Web tools for Documenting, Sharing and Enriching Iraqi Artistic Expressions

Kansa, Sarah Whitcher

skansa@alexandriaarchive.org
The Alexandria Archive Institute

Shabout, Nada

Nada.Shabout@unt.edu
The University of North Texas

Al-Bahloly, Saleem

saleemha@berkeley.edu
University of California, Berkeley

1. Overview

The Modern Art Iraq Archive (MAIA) project is a participatory content-management system to share, trace and enable community enrichment of the modern art heritage of Iraq. The focus of the project is thousands of works of art, many of them now lost, from the Iraqi Museum of Modern Art in Baghdad. MAIA is unique in that it not only documents the lost artworks, but also provides tools for community enhancement of those works, allowing contribution of stories, knowledge and documentation to the system, as well as syndication of the content elsewhere on the web.

For the past eight months, participants in this project have been building a comprehensive virtual archive of the works in the Museum's various galleries, including a database of images and information about the objects (artist name, title, date, dimensions, subject matter, medium, condition, current location, related works, etc). These significant national treasures are displayed in an open format that invites participation from users worldwide, including the Iraqi national and expatriate communities, and users will be encouraged to help identify and understand individual pieces. The MAIA system, which integrates two extant content management systems, Open Context and Omeka, will provide a valuable research tool for scholars, students, as well as the general public, but most importantly for Iraqis: these works of art form an important expression of the Iraqi national experience.

2. Project History

The Iraqi Museum of Modern Art, formerly the Saddam Center for the Arts (Markaz Saddam lil Funun), was established in 1986 as Iraq's museum of modern and contemporary art. During the invasion of Baghdad in April 2003, the structure was severely damaged by fire and looters. Without security and protection from the occupying powers after the collapse of the Baath regime, its collections of approximately 8,000 modern and contemporary Iraqi paintings, sculptures, drawings and photography, dating from late 19th century until April 2003, were entirely looted. Prof. Nada Shabout's research based on sources inside of Iraq indicated that while some works were smuggled outside of the country, most works were still on the market for sale in Baghdad. At an early stage after the invasion, about 1,300 works were found at the National Gallery's basement. They have since been stored at a facility administered by the Ministry of Culture, without restoration, authentication or archiving.

While the fate of the collection is tragic enough, what exasperated the situation further is that the Museum's inventory and documentation disappeared with the works as well, meaning that missing works cannot be traced or repatriated. Nada Shabout has spent the last three years collecting and digitizing all available information about the lost works through meetings with artists, gallery owners, and art educators. In this time, she has found that the situation is dire, with improper documentation and accessioning procedures, scant publication and recording in catalogs, and a lack of inventory for the two decades before the invasion. In the end, the richest available information is in fact in the recollections of individual people; hence, the imperative to develop ways for people to share their knowledge of these works.

3. Approach

The MAIA prototype integrates two existing, open source, content delivery systems, Omeka and Open Context. Omeka (<http://omeka.org/>) is an open source, collections-based publishing platform that allows individuals and organizations to share collections, structure content into exhibits and write essays. It offers customizable themes and a suite of easy-to-install add-ons for customizing site appearance and functionality. Omeka brings Web 2.0 technologies and approaches to academic and cultural websites to foster user interaction and participation, while making design easy with a simple and flexible templating system. Robust open-

source developer and user communities underwrite Omeka's stability and sustainability.¹

Open Context (<http://www.opencontext.org>) is a web-based, open access data publication system that supports enhanced sharing of museum collections and field research data by enabling researchers and cultural heritage collections managers to publish their primary field data, notes and media (images, maps, drawings, videos) on the web. It is free and uses open source software built on common open source technologies (Apache-Solr, MySQL, PHP, and Dojo Ajax) widely accessible and supported by a vast global developer community. Open Context uses Apache-Solr to power a "faceted browse" tool, which allows for much more informed navigation and understanding of collections than the "type and hope" approach of simple key-word searches. This component also delivers web-services, enabling a feed-based approach to syndicating content and integrating collections distributed across the web. These web-services represent a powerful, scalable, and elegantly simple way to facilitate aggregation across multiple collections.

MAIA's approach integrates features of these two systems to maximize the collection's reach, discoverability and creative reuse. Omeka offers a user-friendly platform for building, customizing and organizing the MAIA collection, as well as allowing options for contributions and comments from the community. However, search functionality is limited and Omeka content is largely confined within each Omeka instance, despite its support of OAI/PMH and some feed capabilities. Finally, while Omeka offers stable URLs for every item, Omeka users need to make additional arrangements for archiving their collections. Open Context complements Omeka's capabilities with a faceted browse "plugin," offering powerful web-service capabilities that enable distributed search and syndication of content. These capabilities make Open Context a more powerful platform for supporting aggregation and mashups. Open Context's faceted browse tool provides a much more informed overview of a collection, showing fields associated with content even for custom metadata, allowing exploration beyond simple searches. The Open Context plug in will also greatly expand Omeka's feed capabilities, allowing users to draw custom feeds tailored to their specific interests (such as a particular artist, time period and/or region). Any Omeka site implementing the Open Context faceted browse plugin will be indexed by Open Context, opening up Omeka content to dynamic searching across multiple collections. Finally, Omeka collections using the Open Context plug-in will benefit from accessioning by the California Digital Library. Thus, Omeka's user-friendly collections management and publishing

functions are joined with Open Context's powerful web services to increase the reach and potential for reuse of MAIA content.

Taking advantage of the flexibility in content-management and sharing provided by the integration of these two powerful, open source systems, the MAIA platform is available for free on the web and offers the following additional features:

- **Localization:** All static content in the system is translated into Arabic. Participatory tagging and commentary features are also available in multiple languages.
- **Community input:** A tagging system allows users to comment on any item in the database, thus enriching the content and helping build a memory of the lost works. Users can also link to external content related to the works.
- **Citation:** Easy citation retrieval makes MAIA a useful research tool. Unique citations are generated for every single item in the system, and Omeka expresses bibliographic metadata in a format that the popular Zotero citation management tool can recognize.
- **Copyright Pragmatism:** The creators of many of the works included in the website are unknown, and it is not possible to get permissions from them to disseminate their work online. However, such dissemination is essential if these creators are ever to be identified. Therefore, this project hosts works even in cases where the creator is not known. A clear "take-down" policy ensures that artists can request that their work be removed from the website. In cases where copyright permissions can be obtained, a Creative Commons Attribution, Noncommercial, Share-alike license <http://creativecommons.org/licenses/by-nc-sa/3.0/> is used.

4. Outcomes

By archiving and documenting the known modern artistic works from the Iraqi Modern Art Museum, this project contributes to the preservation of Iraq's cultural heritage. However, the project's greater success lies in the amount of exposure and the quantity and quality of community participation that the project garners. That is, free global exposure of the known content is helpful and informative, but active community input that enriches the works, and perhaps locates some of the lost works, is ideal. We are currently working on maximizing the dissemination of MAIA content, through the use of blogs, interviews, publications and presentations. Our long-term vision for MAIA is that it will become a virtual museum, where visitors will navigate through a map-based interface, exploring galleries

and viewing individual works of art, ideally in the place they stood before the museum was damaged and many of the works lost. In this way, the public will have a visual understanding of the number of works still missing or for which no documentation exists. The emotional impact of seeing blank sections of gallery walls is far greater than reading a number or percentage, and will give the public a more profound understanding of the loss of these works of modern Iraqi heritage. On a more positive note, a visualization of the rich, related content around many of these works will enrich the visitor's experience and understanding of any single work.

References

- Cohen, D. J. and Rosenzweig, R.** (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, Sarah Whitcher, Eric C. Kansa and Ethan Watrall (ed.)**. *Archaeology 2.0 and Beyond: New Tools for Collaboration and Communication*. Submitted to Cotsen Institute of Archaeology Press (In review, Oct 2009).
- Kansa, E. C. and Whitcher Kansa, S. W.** (2009). "Mashable" heritage: formats, licenses and the allure of openness'. *Heritage in the Digital Era*. London: Multi-Science Publishers, pp. 105-112.
- Kansa, S. W. and Kansa, E. C.** (2009). 'Yes, it is all about you: User needs, archaeology and digital data'. *CSA Newsletter*. **22(1)**. <http://csanet.org/newsletter/spring09/nls0902.html>.
- Kansa, S. W. and Kansa, E. C.** (2009). 'Open Context: Developing Common Solutions for Data Sharing'. *CSA Newsletter*. **21(3)**. <http://csanet.org/newsletter/winter09/nlw0902.html>.
- Kansa, S. W. and Kansa, E. C.** (2007). 'Open Content in Open Context'. *Educational Technology Magazine*. **47**: 26-31.
- Kansa, S. W. and Kansa, E. C.** (2007). 'Open Context: Collaborative Data Publication to Bridge Field Research and Museum Collections'. *International Cultural Heritage Informatics Meeting (ICHIM07): Proceedings*. Toronto: Archives & Museum Informatics, 2007. <http://www.archimuse.com/ichim07/papers/kansa/kansa.html>.
- Rosenzweig, R.** (2007). 'Collaboration and the cyberinfrastructure: Academic collaboration with museums and libraries in the digital era'. *First Monday*. **12(7)**.
- Shabout, Nada** (2006). 'The Iraqi Museum of Modern Art: Ethical Implications'. *Collections*. **2(4)**.
- Shabout, Nada** (2006). 'Historiographic Invisibilities: The Case of Contemporary Iraqi Art'. *International Journal of the Humanities*. **3(9)**.
- Shabout, Nada** (2006). 'The "Free" Art of Occupation: Images for a "New" Iraq'. *Arab Studies Quarterly*. **28(3&4)**.
- Shabout, Nada** (2006). 'Preservation of Iraqi Modern Heritage in the Aftermath of the US Invasion of 2003'. *An anthology on Ethics in the Art World*. Levin, G. and King, E. A. (ed.). Allworth Press.

Notes

1. This section has been adapted from: <http://omeka.org/about/>

A Data Model for Digital Musicology and its Current State – The Music Encoding Initiative

Kepper, Johannes

kepper@edirom.de

University of Paderborn

During the last 10 years, XML has gained general acceptance as a data model in the Digital Humanities. Actually, it even leveraged the success of digital projects in the humanities. Meanwhile TEI is the unchallenged standard for all projects in the fields of literature studies, epigraphy, linguistics, history sciences and so on. Many thoughts were invested to bring TEI and other related formats like EpiDoc to a level that suffices general scholarly needs.

At first sight, things went differently in the field of music encoding. Around the year 2000 a couple of XML-based encoding schemes for music notation emerged, and within just a few years MusicXML became the best-known and most widespread music encoding format. It was intended to serve as an interchange format between different music applications, and even today it is virtually indispensable for this very important task. At the same time, this orientation of MusicXML requires a certain "simplicity" that facilitates implementations in various applications.

The Music Encoding Initiative (MEI) went a different way. Not aiming at application support in the first place, an encoding model for *scholarly* purposes was developed over years. Strongly influenced by the concepts of the TEI, Perry Roland as initiator of the format tried to transfer these concepts to the field of music encoding. For large parts, this is quite easy: music notation is a kind of text, and many unspecific modules of TEI can be reused for music encoding with only small changes. But then again, music notation itself offers a much higher complexity than other texts. It is multi-dimensional not only because of its layout of multiple vertically aligned staves, but also because of its simultaneity of harmonic and melodic progression. In music notation, the text itself consists of overlapping hierarchies and therefore demands a quite sophisticated data model. Most often, it is virtually impossible to preserve all possible meanings (or better: interpretations) of a musical text with reasonable effort. The reason is that the written text is only a part of the complete information. Every notation serves a certain purpose, and each composer or copyist uses

only as many symbols as he needs to be explicit to his contemporaries. Besides this, the "rules" of music notation changed significantly over time, even though these developments often seem to be very subtle.

All this leads to the problem that there is no absolutely fixed terminology in music notation. Some phenomena are still not completely understood or even defined, such as the problem of dots, strokes and hooks in scores from the classical period. The lack of a complete and well-defined terminology even for restricted repertoires makes the encoding of music notation on a scholarly level highly demanding, and, at the same time, the implementation and usage of such an encoding scheme is anything but trivial.

The Music Encoding Initiative has chosen this way, and currently it stands on an important turning point: In a one-year project funded by the NEH and DFG the original model was revised and has proven to meet all essential scholarly requirements for such a format. In the next years, it needs to be disseminated in the fields of musicology, music information retrieval, music philology and digital humanities in general. A first step in this direction is the TEI's Special Interest Group on music encoding, whose members were actively involved in the recent developments on MEI, and who seek to find ways to bring MEI and TEI closer together.

Due to the complexity of music notation – and thus music notation encoding too – application support for MEI is crucial to ensure its dissemination: Almost no traditional musicologist would be willing to work with a XML-editor like Oxygen. There are several projects currently working on such applications for MEI: The DiMusEd-Project, situated in Tübingen (Germany), uses SVG to render encodings of multiple sources of music notated with medieval neumes. Although this repertoire uses a limited set of symbols, this project already shows the benefits of a dynamic rendering from an encoding instead of engraved scores. The Edirom project, (Detmold, Germany) aims to establish workflows for digital scholarly editions of music. In the application for preparing such editions it is already using MEI to store all structural information about the musical text as well as the containing documents. For moving from basically facsimile-based editions to completely digital editions it is planned to offer complete encodings of all relevant sources including the rendering-facilities already demonstrated by the DiMusEd-project. In order to achieve this goal Edirom closely collaborates with the most ambitious of all ongoing MEI-related projects: TextGrid. A sub-project of this major German initiative, which is also located in Detmold, seeks to develop a limited scorewriter for MEI offering a graphical user interface for musicologists. In this case „limited“

means that the project neither intends to support MEI completely nor tries to keep up with the engraving quality of already existing scorewriters: the unambiguity of the output is more important than its beauty.

All these German projects collaborate closely with the ongoing efforts in the US to further improve the format itself and to provide interchange to other relevant formats such as Humdrum and MusicXML. Depending on further funding by NEH and DFG respectively it is intended to provide reasonable collections of MEI encodings to facilitate further usage of the format. Although MEI will not find the wide acceptance MusicXML already has, all these components will help to disseminate MEI in the academic world, to promote interchange of high-quality data and to explore new methods for digital representations of written music.

The talk will provide a short introduction to the current state of MEI – both the format itself and the projects and applications already working on and with it.

From Text to Image to Analysis: Visualization of Chinese Buddhist Canon

Lancaster, Lewis

buddhst@berkeley.edu

University of California, Berkeley

This presentation is based on software interface development by a team at the University of California, Berkeley. The database which was used for this technology is the digital version of the Korean Buddhist canon written in Chinese characters. The tool shown was built with the help of a two year grant of support (2007-2009) from the National Science Foundation. International collaboration has included the Institute of Tripitaka Koreana in Seoul who provided scanned images of rubbings taken from the original printing blocks at Hae-in Monastery. The software metadata is based on the previous publication *The Korean Buddhist Canon: A Descriptive Catalogue* (Lancaster, 1979).¹ A digital version of this catalogue was made by Charles Muller of Tokyo University who has made it freely available on the internet (Muller, 2004).² The project has been a part of the Electronic Cultural Atlas Initiative (ECAI) and received support from that group's *Atlas of Chinese Religions* research funded by the Luce Foundation. This atlas is being constructed in collaboration with the GIS Center at Academia Sinica in Taiwan and will provide references to the place names associated with the production of the translations and compilations included in the canon. Continued research on developing the software is being done in cooperation with the School of Creative Media and the Department of Chinese Translations Linguistics at City University of Hong Kong. It is important to understand that no project of this kind could possibly be undertaken without these multiple and widespread collaborations.

In the example being described in this presentation, we use the software to focus on the digital version of the 13th century Korean printing block edition of the Buddhist canon (Lancaster, 1996).³ The canon, represented on blocks, contains more than 52 million characters/glyphs carved onto 166,000 surfaces each producing a page of text when printed. The number of lines, containing up to 14 glyphs, on the plates number over three million. The entire set of the canon is divided into 1,514 different texts representing dated translations and compilations made over a period of seven centuries. The size of the data, the temporal span of its composition, and the history of

acquisition in Korea of the hundreds of texts from China, provide us with a reasonable challenge for the interface design.

The previous approach to the study of this canon was the traditional analytical one of close reading of specific examples of texts followed by a search through a defined corpus for additional examples. When confronted with 166,000 pages, such activity had to be limited. As a result, analysis was made without having a full picture of the use of target words throughout the entire collection of texts. That is to say, our scholarship was often determined and limited by externalities such as availability, access, and size of written material. In order to overcome these problems, scholars tended to seek for a reduced body of material that was deemed to be important by the weight of academic precedent.

In the current digital age, however, the limits on “what can be considered” in the Korean Buddhist canon have been significantly removed. We can consider all of the texts, all of the words, and all of the metadata in every search. Consequently, the practices of traditional scholarship for the canon have begun to falter. When the entire canon had been digitized in the last decade of the 20th century, the process of search and retrieval of target words and phrases was transformed. Nonetheless, problems remain for Buddhist scholars using this digital version. In many cases, the menu which appears after a search of a term can contain thousands of references. The references presented as a display of each line where the word occurs can still occupy long hours of time to analyze and put into some form of presentation.

We are in need of new ways to display search results that will allow scholars to quickly perceive such things as the patterns of occurrences, examples of clustering, view of target words with adjacent companion words, graphic models of profiles of sequence, computation of occurrences not only for the whole of the set but also broken down by text and date. The displays give the researcher aids in evaluating and analyzing the patterns for each word.

As a first step, we look for the number of times that each of the characters appears in the canon. As each search is made, a report appears in visual form on a “ribbon” of blue dots, where each of the dots represents one of the 52 million glyphs. The “ribbon” is more than a picture for each “blue dot” has 35 fields of metadata behind it. It is marked for exact placement on the original printing block, date of translation, name of translator for the text in which it appears, UNICODE number, place of translation, name of text containing the example, etc. The dots are arranged by “panes” that correspond to the more than 160,000 pages of the version preserved

in Korea. The dot is an abstract image that permits the user to see patterns of occurrence without the barrier of complex display of natural language glyph constructions (Lancaster, 2009).⁴

It is at this first step that we note the distinct shift in methodology. The initial move on the part of the scholar, who uses this interface, is to turn directly to the data itself rather than to reference works. This is accomplished because the software provides a process of searching through the entire set of data at once. There is no step of consulting a reference work such as a concordance before proceeding to the text itself.

This visual becomes the first factor in the scholar’s “work flow” planning. It shows whether the glyph(s) are scattered throughout the canon, whether there are heavy concentrations in a few places, whether there are only a few examples that appear in widely separated examples in terms of texts, time, and translators. Securing this much information within a few seconds can be compared to the hours of effort it would take to construct such an analysis of patterning, even with an internet search for each example of the term. In every case, the “blue ribbon” and other graphics are displaying a large amount of data in the visual form. We can “see” the occurrences of our target search within the 52 million glyphs and immediately understand the nature of the patterning.

Words in the canon have a history and we can begin to spot the ways in which they evolve. Since each dot has multiple metadata fields, they need not be seen only in the sequence of the canonic arrangement. They can be rearranged by time of translation, by translator, or place of translation. The visuals will quickly show that some words grow in number of occurrences over time and others fade into a more marginal role.

There have been surprises from the use of the tool on the canonic words. The computation for an important expression can help us identify apocryphal texts, or texts that show the characteristics of translation rather than compilation. “Companion” words that become associated with a target word can be used to identify the primary meaning of an occurrence.

Future plans call for the exploration of making this tool multi-lingual and provision for having it available as an open source and free add-on to datasets.

Funding

- This material is based upon work supported by the National Science Foundation under Grant No. 0840061.
- Support for the Atlas of Chinese Religions given by grants from the Henry Luce Foundation, Inc. New York.

- Support for Atlas of Chinese Religions provided in part by the GIS Center of Academia Sinica, Taiwan.
-

References

- Lancaster, Lewis, Park, Sungbae** (1979). *The Korean Buddhist Canon: A Descriptive Catalogue*. Berkeley: Berkeley University Press.
- Lancaster, Lewis** (1996). 'The Buddhist Canon in the Koryo Period'. *Buddhism in Koryo: A Royal Religion*. Kikun Sub, Chai-shin Yu (eds.). Berkeley: Berkeley University Press.
- Lancaster, Lewis** (2009). *SGER: Text Analysis and Pattern Detection: 3-D and Virtual Reality Environments*. <http://ecai.org/textpatternanalysis/>.
- Muller, Charles** (2004). *Digital Version of The Korean Buddhist Canon: A Descriptive Catalogue*. http://www.acmuller.net/descriptive_catalogue/index.html.

Notes

1. Lewis Lancaster and Sungbae Park. *The Korean Buddhist Canon: A Descriptive Catalogue*. Berkeley: University Press. 1979.
2. See http://www.acmuller.net/descriptive_catalogue/index.html for the digital version of the *The Korean Buddhist Canon: A Descriptive Catalogue* on his server in Tokyo. (2004).
3. A description of this canon is found in my article "The Buddhist Canon in the Koryo Period," *Buddhism in Koryo: A Royal Religion*, edited by Kikun Suh and Chai-shin Yu. Korea Research Monograph #21, Institute of East Asian Studies, University of California: Berkeley, 1996.
4. See examples of the interface and report of progress in my report to NSF: *SGER: Text Analysis and Pattern Detection: 3-D and Virtual Reality Environments* (2009). <http://ecai.org/textpatternanalysis/>

Crossing the Boundary: Exploring the Educational Potential of Social Networking Sites

Lang, Anouk

a.lang@qmul.ac.uk

Queen Mary University of London

To date, the scholarship on social networking sites (SNSs) such as Facebook and MySpace has focused largely on areas other than pedagogy, with Boyd and Ellison (2007) observing that so far, most SNS research has centred on "impression management and friendship performance, networks and network structure, online/offline connections, and privacy issues". Some work has been done on the effect of instructor presence on Facebook (Hewitt and Forte, 2006; Mazer et al., 2007; Szwelnik, 2008), the creation of MySpace pages in terms of the acquisition of new forms of digital literacy (Perkel, 2008), and some of the difficulties and benefits of SNSs for university students (Thelwall, 2008). Aside from this, however, there is little scholarship on the educational uses and potential of SNSs at university level. As Szwelnik (2008) comments, 'Facebook has attracted a lot of attention from media and business but not yet a lot of attention from educational researchers'.

It is perhaps not surprising that it is the social aspects of these sites that have attracted the most critical attention, given that their central purpose is understood to be the management and navigation of (often pre-existing) relationships, rather than a means by which to share interests, complete tasks, or simply communicate with others (Boyd and Ellison, 2007; OFCOM, 2008). However, given that the social dimension of education is a fundamental to learning, it is worth exploring how SNSs may be used to pedagogical advantage. This is particularly the case given the large proportion of university students that access the sites: Ellison et al. (2007) found that 94% of the undergraduate population at Michigan State University were members, while a 2007 Ipsos Mori poll found that 95% of British undergraduates are regular users (Shepherd, 2008). SNSs, it would seem, are a resource not yet being used to their full potential for university teaching.

This paper reports on the findings of a research project designed to address the under-utilisation of SNSs at university level and the corresponding gap in the literature. Undertaken with students in the School of Languages, Linguistics and Film at

Queen Mary, University of London, it investigates the educational potential of Facebook for facilitating informal learning with students on their year abroad, particularly in the domain of intercultural awareness and communication. Events held in previous years demonstrated the usefulness of bringing language students together in face-to-face contexts to reflect on their own and others' diverse experiences of the year abroad. This project set up an online space on an SNS to facilitate this kind of learning through a peer mentoring framework, and to allow discussions of this sort to occur regularly during the students' time abroad, rather than after it was over.

Undergraduate students in their second and final year of a language course were surveyed about their attitude towards the year abroad, their use of technology and SNSs, and, following Ellison et al. (2007), their affective investment in SNSs. Two different populations were surveyed: second year students organising their year abroad for the following year, and final year students who had returned from their year abroad. The results were used to develop focus group protocols to gauge students' receptivity to the idea of using Facebook to carry out course-related discussions, to judge the extent to which Facebook was a hospitable environment for peer mentoring to occur, and to determine which Facebook applications would best assist with educational objectives. Students were also asked about which aspects of the year abroad they were already using Facebook to engage with, and which elements of their time away could be ameliorated through provision of a virtual meeting place for discussion. Several peer mentors were then chosen from the cohort who had already completed a year abroad, and these students were trained in online moderating and mentoring. A Facebook group was set up for the student mentors to use, and this was observed over a period of three months. Following this, four methods of evaluation were used to measure the effectiveness of the Facebook group: a) an online survey for third-year students currently on their year abroad; b) informal discussions with academic staff in modern languages; c) interviews with the peer mentors held at a computer; and d) close analysis and corpus analysis of the text of the online discussions.

Paul and Brier (2001) and Cummings et al. (2006) have explored how students of university age use Facebook and other internet technologies to alleviate the "friendsickness" brought about by moving away from one's friends. This project aimed to capitalise on the powerful ability of SNSs to address this relational need by drawing students into online conversations and collaborations that not only helped them to sustain relationships but also to use those relationships to learn from one another. However,

existing research points to a strong resistance from university students to academics occupying "their" space on an SNS, something Szwelnik terms "crossing the boundary" (2008). Hewitt and Forte (2003) observe that identity management is a significant concern for SNS users when the roles they occupy cross perceived social boundaries and bring organizational power relationships into visibility, citing one student's fears that Facebook could "unfairly skew a professor's perception of a student in a student environment". Given that both social boundaries and uneven power relationships both come into play in the context of teacher-led discussions around course-related material, the project sought to find a way to build a learning community without infringing on a space perceived not to belong to academic staff, and to shift the discursive content from social to educational without forcing students to "cross the boundary". In working with peer mentors, the project aimed not only to avoid these boundary-crossing problems but also to work intentionality into the fabric of the Facebook group. As Woods and Ebersole (2003) observe, transforming textual exchanges into a learning community with a positive social dynamic requires intentional decisions in the realm of both verbal and nonverbal communication, so student mentors needed to be made acquainted with the learning objectives for the educational context, and carefully trained in techniques of e-moderation to overcome the challenges a mediated environment can pose to productive discussions. This approach also had the advantage of being, to an extent, futureproofed: students were more likely than academic staff to know which technologies are most popular with their peers, and once a framework for online mentoring is established, this could be moved in future years to different sites or applications as students' usage patterns change. A further advantage of this model is that it equips the student mentors with the digital literacy and communication skills to operate in the kinds of virtual environments that, as knowledge workers, they are likely to inhabit in their careers.

In summarising the results of this study, this paper will report on the benefits of using an SNS to support informal learning in the context of students on a year abroad, and set out approaches that universities can take to promote learning more generally through the use of SNSs. It uses tools from the Internet Community Text Analyzer (<http://textanalytics.net/>) to visualise the networks that developed between students on the site, and to identify productive models of SNS-mediated student mentoring behaviour.

References

- Boyd, D. M., Ellison, N. B.** (2007). 'Social Network Sites: Definition, History, and Scholarship'. *Journal of Computer-Mediated Communication*. **13(1)**: 210-230. <http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html> (accessed 19 September 2008).
- Cummings, J., Lee, J., Kraut, R.** (2006). 'Communication Technology and Friendship During the Transition from High School to college'. *Computers, Phones, and the Internet: Domesticating Information Technology*. Kraut, R. E., Brynin, M., Kiesler, S. (eds.). New York: Oxford University Press, pp. pp. 265-278.
- Ellison, N. B., Steinfield, C., Lampe, C.** (2007). 'The Benefits of Facebook "Friends": Social Capital and College Students' Use of Online Social Network Sites'. *Journal of Computer-Mediated Communication*. **12(4)**: 1143-1168. <http://www3.interscience.wiley.com/cgi-bin/fulltext/117979349/HTMLSTART> (accessed 30 October 2008).
- Hewitt, A., Forte, A.** (2006). 'Crossing Boundaries: Identity Management and Student/Faculty Relationships on the Facebook'. CSCW'06. Banff, AB, November 2006. <http://www-static.cc.gatech.edu/~aforte/HewittForteCSCWPoster2006.pdf> (accessed 19 September 2008).
- Mazer, J., Murphy, R., Simonds, C.** (2007). 'I'll See You On "Facebook": The Effects of Computer-Mediated Teacher Self-Disclosure on Student Motivation, Affective Learning, and Classroom Climate'. *Communication Education*. **56(1)**: 1-17.
- OFCOM [Office of Communication]** (2008). *Social Networking: A Quantitative and Qualitative Research Report into Attitudes, Behaviours and Use*. http://www.ofcom.org.uk/advice/media_literacy/medlitpub/medlitpubrss/socialnetworking/report.pdf (accessed 22 September 2008).
- Paul, E. L., Brier, S.** (2001). 'Friendsickness in the Transition to College: Precollege Predictors and College Adjustment Correlates'. *Journal of Counseling & Development*. **79(1)**: 77-89.
- Perkel, D.** (2008). 'Copy and Paste Literacy? Literacy Practices in the Production of a MySpace Profile'. *Informal Learning and Digital Media: Constructions, Contexts, Consequences*. Drotner, K., Jensen, H.S., Schroeder, K. (eds.). Newcastle, UK: Cambridge Scholars Press, pp. 203-224.
- Shepherd, J.** (30 September 2008). 'Make Friends Before you Start'. *The Guardian*. <http://www.guardian.co.uk/education/2008/sep/30/students.facebook>.
- Szwelnik, A.** (2008) (11 November 2008). *BMAF Project Report: Embracing the Web 2.0 Culture in Business Education – The New Face of Facebook*. http://www.heacademy.ac.uk/assets/bmaf/documents/projects/TRDG_projects/trdg_0708/finalreports_0708/Alice_Szwelnik_OBU_web.doc.
- Thelwall, M.** (2008) (25 January 2008). 'MySpace, Facebook, Bebo: Social Networking Students'. *Association for Learning Technology Online Newsletter*. **11**. http://newsletter.alt.ac.uk/e_article000993849.cfm.
- Woods, R., Ebersole, S.** (2003). 'Using Non-Subject-Matter-Specific Discussion Boards to Build Connectedness in Online Learning'. *American Journal of Distance Education*. **17(2)**: 99-118.

Queste del Saint Graal: Textometry Platform on the Service of a Scholarly Edition

Lavrentiev, Alexei

Alexei.Lavrentev@ens-lyon.fr
UMR ICAR Université de Lyon / CNRS

Serge, Heiden

slh@ens-lsh.fr
UMR ICAR Université de Lyon / CNRS

Yepdieu, Adrien

Politecnico di Torino, student

In this poster/demo we will present an online scholarly edition of *La queste del saint Graal* (*The Quest for the Holy Grail*) based on a manuscript of Lyons public library (Lyon, BM, P.A. 77) and built in the Textometry platform (TXM). The particularity of this edition is that it combines rich paleographic and philological data (including digital photographs, various layers of transcription, translation in modern French and editorial notes) with advanced linguistic search and analysis tools provided by textometric software.

Despite some damage (torn miniature in the beginning, missing folios in the end), Lyons manuscript is considered to be one of the best witnesses of the *Queste del saint Graal*, a 13th century French novel, part of the famous "Arthurian" prose cycle. A well-known edition by Albert Pauphilet based on this manuscript was first published in 1923 and has been regularly reprinted ever since. However, this edition cannot be used as a trustworthy source of linguistic data, as it contains multiple corrections, and the readings of the primary source are not always accessible.

In the late 1990s Christiane Marchello-Nizia started working on a new edition, which was supposed to be closer to the witness, explicitly correcting only doubtless scribal errors and equipped with multiple tools to assist the reader in understanding the text and to explore various features of its language. This work was to a large extent inspired by the experience of the *Charrette* Project (Pignatelli & Robinson 2002). The first prototype of the edition was published on the web in 2002, and a completely new version was released in 2009. The edition is still under construction but its main components are already in place and can be accessed on the Web through an early TXM prototype.

These are the following:

- Old French text (108 000 words);
- digital facsimile of the manuscript (418 text columns);
- modern French translation;
- scholarly introduction;
- tools for textometric analysis.

The Old French text is established according to rigorous editorial principles stated in the Introduction. A particular feature of the edition is the respect of scribal punctuation. It is encoded according to the XML-TEI standard with special attention to linguistic data. All words are explicitly tagged and annotated for part of speech using the Cattex2009 tagset (Prévost&al, to be published). The text can be displayed and searched in several presentation forms: a "traditional" view (*vue courante*), close to the norms of French critical editions (Vielliard & Guyotjeannin 2001), a "diplomatic" view respecting linguistically significant graphical features of the manuscript (such as the absence of "phonetic" distinction between *u* and *v*), and a "facsimile" view where all noticeable graphical distinctions of the primary source are represented. For instance, medieval abbreviations are tacitly expanded in the traditional view, they are expanded but typographically highlighted (using italics) in the diplomatic view and they are represented by abbreviation marks in the facsimile view. The concept of three different views of the source text is based on the multi-level transcription model of the Medieval Nordic Text Archive (Haugen 2004). MUFI character codes (Haugen 2009) are used for "special" medieval characters in the facsimile view (such as abbreviation marks or letter variants). A MUFI compliant font (such as Andron Scriptor Web) needs to be installed on the system to display these characters correctly.

The three views of the Old French text, the translation and the photographs can be browsed individually or side-by-side in two columns. The 'diplomatic' and 'facsimile' views, as well as the translation, are only available for a few folios at present but they will be progressively edited for the whole text.

The innovative aspect of this new edition consists first of all in the integration of textometric research tools. Those tools assist the reader by offering qualitative services like full text search engine to build KWIC concordances and browse the text through them, and quantitative services like comparing statistically the occurrences of different linguistic phenomena in various parts of the text or analyzing statistically their collocational attraction inside various contexts, such as sentences.

One can open the textometric panel by clicking on the *Outils* ('tools') button in the bottom of the window. The platform makes it possible to search for linguistic data (words, parts of words, parts of speech, phrases, etc.) in any presentation form using the powerful CQP query language (IMS Open Corpus Workbench) and to display KWIC concordances of the search matches with customizable context size and sort options. Concordances are displayed in a new tab next to the search form. The corresponding page (or column) of the edition is displayed upon a click on a line corresponding to an occurrence in the concordance. The words matching the query are then highlighted (see the figure below).

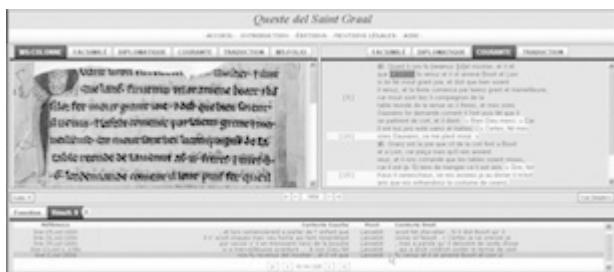


Figure 1

In the future, all the main textometric tools will be interfaced in the edition. These include specificity, collocates, lexicograms, etc. The source XML-TEI files of the Old French text and its modern French translation will be downloadable under a Creative Commons license (Attribution-Noncommercial-Share Alike) as soon as the edition is stable. The editorial and online textometric research platform is distributed with an open source license (GPL v3) and can be used for publishing on the web various texts with XML-TEI encoding. The platform can handle any number of texts, up to several hundred million words in total.

The poster will display the editorial principles and several screenshots of the edition. Interactive work with the edition can be performed at a demo session.

References

Haugen, Odd E. (2004). 'Parallel Views: Multi-level Encoding of Medieval Nordic Primary Sources'. *Literary and Linguistic Computing*. **19.1**: 73-91.

Haugen, Odd E., ed. (2009). *MUFI character recommendation. Characters in the official Unicode Standard and in the Private Use Area for Medieval texts written in the Latin alphabet. Part 1: Alphabetical order. Version 3.0*. Bergen: Medieval Unicode Font Initiative.

Pauphilet, Albert. (1923). *La Queste del Saint Graal. Roman du XIII^e siècle*. Paris: Champion.

Pignatelli, Cinzia and Molly Robinson, eds. (2002). *Chrétien de Troyes : Le chevalier de la Charette (Lancelot) : le "Projet Charette" et le renouvellement de la critique philologique des textes, Œuvres et critique*, 27.1. Tübingen: Gunter Narr Verlag.

Prévost, Sophie, Céline Guillot, Alexei Lavrentiev and Serge Heiden (To be published). *Jeu d'étiquettes CATTEX2009*. Lyon: Équipe de la BFM. <http://ccfm.ens-lsh.fr>.

Vielliard, Françoise and Olivier Guyotjeannin (2001). *Conseils pour l'édition des textes médiévaux, Fascicule I, Conseils généraux*. Paris: CTHS, Ecole nationale des chartes.

Andron Scriptor font page. <http://www.mufl.info/fonts/#Andron>.

Charrette Project. <http://lancelot.baylor.edu>.

Creative Commons license. <http://creativecommons.org/licenses/by-nc-sa/3.0>.

IMS Open Corpus Workbench. <http://cwb.sourceforge.net>.

Medieval Unicode Font Initiative (MUFI). <http://www.mufl.info>.

Queste del Saint Graal edition prototype (temporary). <http://textometrie.risc.cnrs.fr/txm>.

Textometry Project. <http://textometrie.ens-lsh.fr>.

TXM source code. <https://sourceforge.net/projects/textometrie>.

The Graceful Degradation Survey: Managing Digital Humanities Projects Through Times of Transition and Decline

Nowviskie, Bethany

bethany@virginia.edu

Scholars' Lab, University of Virginia Library

Porter, Dot

dot.porter@gmail.com

Digital Humanities Observatory, Royal Irish Academy

Transition and decline are pressing issues for scholars in the digital humanities, as our projects tend to be both collaborative and open-ended. Project staff relocate, reestablish themselves in new areas, or retire, even as funding and institutional support comes and goes. How are projects to be designed so that they can be maintained, or maintain themselves, through periods of change? How might projects be designed in a way that takes periods of transition and possible decline into account from the very start?

These are some of the issues we sought to explore in undertaking "Graceful Degradation: Managing Digital Projects in Times of Transition and Decline," a wide-ranging survey of the digital humanities community, in the summer of 2009. Our intent was to investigate how the community currently deals with these problems and, using our survey data – which also included some demographic information and measures of perceived levels of support and impact of various kinds of change – to make recommendations on how we, as a community, might improve the current approach.

This presentation will provide a detailed look at the outcomes of the "Graceful Degradation" survey, and propose some initial recommendations. (Full recommendations will be published in a separate article.)

The survey was designed in consultation with statistical analysis staff at the Scholars' Lab, University of Virginia Library and unveiled at Digital Humanities 2009 in College Park, Maryland and at Digital Resources for the Humanities and Arts (DRHA 2009) in Belfast, Northern Ireland. It was conducted online between July and September 2009. There were 102 completed surveys, representing 114 discrete projects. Some of our findings are presented below.

The vast majority (76%) of Graceful Degradation respondents come from "large universities with a research emphasis," but teaching colleges, cultural heritage institutions, and commercial ventures were also represented. Most respondents have worked in project management or digital research and development efforts in the humanities for 2-10 years, but 35% of respondents have been engaged in this activity for more than a decade.

Respondents were asked to rate perceived levels of support for the digital humanities at their home institutions, including (as separate queries) general support, support for collaborative activities, local funding and cost-share opportunities, support by higher administration, department-level or local support, and support for project management and grant-writing.

64% of respondents had experienced the decline of a project or had weathered a period of difficult transition. 29% of respondents indicated a sense that digital humanities projects are more likely to decline or suffer these difficult transitions at their institutions than at others.

Participants were asked to respond in detail regarding their experiences with a particular project that suffered decline or a difficult transition. The following percentages apply to the primary or to the single project which survey participants addressed. 37% of respondents identified themselves as project lead or principal investigator (PI) for the project they discussed in depth. 29% of respondents self-identified as project managers, and other respondents fell into categories such as "dedicated, project-specific support staff," "support staff on loan from other units," "graduate or undergraduate research staff," "post-docs or faculty collaborators."

38% of projects discussed fell into the category of "content creation, digitization, and archive-building," but other categories (including software development, online community-building activities, online journals and other publications, and creation of support infrastructure for digital scholarship) were also represented. Predominant disciplines and time periods addressed were literary and textual studies and digital history, from the modern or early modern era. More projects (31% and 24% respectively) identified an academic department and a library or museum as their primary institutional home, with 23% primarily housed in a digital humanities center.

Of projects that had experienced decline or difficult transition, most were identified as still "ongoing and active" (51%), with 26% abandoned or dormant, and 15% and 8% either complete or "just getting started," respectively.

Participants were asked about funding sources for these projects (generally via institutional support or "external public funding") and understood length of funding or support. Projects treated were generally funded for 2-3 years, with no possibility of renewal, but often (in 21% of cases) the length of funding or support was "unclear." That said, 75% of respondents considered their project's funding to be "reliable and clear in scope."

Most respondents undertook the treated project with clear plans for supporting it beyond an initial funding period, but most projects also ultimately "differed in scope or definition from early plans." In 68% of cases, participants had identified both short-term and long-term goals for their projects, but conscious use of "specific project management techniques or tools" and "risk management strategies" was a rarity. Anecdotal responses treated the impact of varying levels of planning or lack of planning on digital projects.

The majority of projects (55%) experienced no negative impact due to staff overturn whatsoever. For projects that did, we asked participants to rate the negative impact of overturn of six different categories of staff members and collaborators. Survey participants also rated the broad impact of their projects in a dozen areas, such as "scholarly inquiry in a particular field," "my own pedagogical practice," and "the professional advancement of my collaborators."

Participants were additionally given the opportunity to respond to several prose prompts, and to add more contextual information to many of the questions for which we had devised statistical measures. They summarized the reasons for the project decline or difficult transitions they experienced, and offered formulae for their successes. Some respondents identified nuanced issues with intellectual property and open source as contributing factors. We plan to summarize these rich responses and reveal the results of qualitative data analysis at the conference.

67% of respondents indicated that their personal views and practices have evolved as a result of experiencing a period of difficult transition or the decline of digital humanities project, but in only 32% of cases did they feel that the views or practices of their local institutions or the larger academic community have evolved in response to such experiences like these.

67% of respondents also indicated that they had experienced what they would consider a "phase of successful transition" in their digital projects, and offered anecdotal advice as to what made that possible.

At Digital Humanities 2010, we will summarize and offer some visualizations and analysis of these findings and others, and we will address the extensive qualitative data that were collected from participants in free-form text responses. (Several participants granted us permission to quote their responses directly. We will anonymize and summarize responses from others.) We will also draw conclusions about avenues for future research and – more importantly – identify areas for future action on the part of institutions supporting digital humanities projects and professional societies representing the digital humanities community.

References

- Matthew Kirschenbaum et al.** 'Done: 'Finished' Projects in the Digital Humanities'. *Digital Humanities*. Urbana-Champaign, Illinois, 2007.
- Daniel Pitti** (2004). 'Designing Sustainable Projects and Publications'. *A Companion to Digital Humanities*. Susan Schreibman, Ray Siemens, John Unsworth (eds.). Oxford: Blackwell. <http://digitalhumanities.org/companion/> (accessed 13 November 2009).
- Stephen Ramsay et al.** 'Innovations in Interdisciplinary Research Project Management'. *Digital Humanities*. Oulu, Finland, 2008.
- Geoffrey Rockwell, Shawn Day**. 'Burying Dead Projects: Depositing the Globalization Compendium'. *Digital Humanities*. College Park, Maryland, 2009.
- Siemens, Lynne (ed.)**. *Issues in Large Project Planning and Management*. *Digital Humanities Summer Institute*. Victoria, 2009.
- Siemens, Lynne** (June 1, 2009). 'It's a team if you use "reply all": An exploration of research teams in digital humanities environments'. *Literary and Linguistic Computing Advance Access*. *Literary and Linguistic Computing* 24: 225-233. <http://DOI 10.1093/linc/fqp009>
- Siemens, L., Cunningham, R., Duff, W., Warwick, C.** (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*.
- Diane Zorich** (November 2008.). *A Survey of Digital Humanities Centers in the United States* CLIR Reports, <http://www.clir.org/pubs/reports/pub143/contents.html>.

LAP, LICHEN, and DASS – Experiences combining data and tools

Opas-Hänninen, Lisa Lena

lisa.lena.opas-hanninen@oulu.fi

University of Oulu, Finland

Juuso, Ilkka

ijuuso@ee.oulu.fi

University of Oulu, Finland

Kretzschmar, William A. Jr.

kretzsch@uga.edu

University of Georgia, USA

Seppänen, Tapio

tapio@ee.oulu.fi

University of Oulu, Finland

The Linguistic Atlas team at the University of Georgia (LAP) and the LICHEN research team at the University of Oulu have been investigating the application of advances in information engineering to humanities scholarship, in particular methods for managing and mining large-scale linguistic databases. Our aim has been to bring together the linguistic and technological expertise from Oulu and Georgia in order to develop practical solutions for common problems. In this paper we will show the results of this cooperation--including the Digital Archive of Southern Speech (DASS), the pilot product for LAP-LICHEN released in 2009--and discuss our experiences and lessons learned.

The LAP audio archive, amounting to 7000 hours of interviews, is an unparalleled resource for study not only of the common language of the US but for its culture more generally, stories of daily life in America. Study of LAP interviews so far has taken advantage only of small bits of transcribed data extracted from the full interview. The large, untranscribed bulk of LAP interviews consists of the speakers' accounts of their lives and their families, their occupations and their diversions, their houses and their land. Along with direct questioning and conversational passages, in a quarter of the thousands of audio files so far processed these stories take the form of narratives of at least one minute of continuous speech. To preserve and to make this audio archive available puts the people back into what has seemed to some scholars to be a dry academic exercise of collecting words. The LAP team has always appreciated the personal, individual nature of each interview, as well as the way that the interviews can represent American culture; with

DASS, we can now share that appreciation much more broadly with both the academic community and with the public.

DASS is a collection of 64 interviews from the Linguistic Atlas of the Gulf States (LAGS) selected by LAGS Director Lee Pederson. Four interviews come from each of the sixteen regional LAGS sectors. Within each sector there is one speaker from each Atlas Type: folk (largely uneducated and insular), common (moderate education and experience), cultivated (higher education and/or participation in high culture). One African American speaker was selected from each sector, and folk, common, and cultivated African American speakers are distributed across the sectors. Speakers cover a wide range of ages and social circumstances. Over 400 hours of audio files are provided both as large uncompressed .wav files (useful for acoustic phonetic processing) and as thousands of small .mp3 files for general listening. Files are indexed according to subject matter and speaker, according to a set list of 40 topics. Metadata and finding aids for particular topics and kinds of speakers are provided, including search tools and a GIS function. Together the DASS data and the LICHEN tools comprise about 200 GB of data, provided on a portable USB drive. The interviews were digitized and processed by the LAP team at the University of Georgia with assistance from a grant from the National Endowment for the Humanities (PW-50007, "Digitization of Atlas Audio Recordings", with Opas-Hänninen as partner). The first phase of the LICHEN project was lead jointly by Opas-Hänninen and Seppänen and funded by a grant from the Emil Aaltonen Foundation (2006-2008, with Kretzschmar as an international collaborator). The University of Oulu and the University of Georgia drew up legal agreements regarding copyrights and the distribution of the software with the data. The DASS/LICHEN package is distributed by the LAP.

DASS is only the beginning, however. The research team at the University of Oulu has developed LICHEN as an electronic framework, i.e. a type of toolbox, which handles multimodal data. The toolbox has been developed using two sets of data as testbeds, namely the Oulu Archive of Minority Languages in the North containing samples from the Kven, Meänkieli, Veps and Karelian languages, as well as DASS. Some of the data from minority languages exists as video, which the toolbox handles along with audio and text. We are now working on a transcription tool, so that audio and video materials can be provided with textual representations aligned with the sound and video. Finally, we are rebuilding LICHEN as a Web-enabled framework, so that users can access our language and cultural materials remotely in line with the movement for the creation of public corpora (see, e.g., Kretzschmar, Anderson,

Beal, Corrigan, Opas-Hänninen, and Plichta 2006; Kretzschmar, Childs, and Dollinger 2009).

LAP-LICHEN cooperation began in 2004, when Kretzschmar, Opas-Hänninen, Anderson, Beal, and Corrigan met in Newcastle, to follow up on conversations about public corpora begun earlier. The group found that there were common problems, standards, best practices for the corpora managed by those attending, and agreed to prepare a presentation at ICAME in 2005 to highlight the possibility for shared methods and joint actions (later published as the 2006 programmatic article). The LAP-LICHEN collaboration bloomed as a result. Grants were obtained for cooperation: the LICHEN model was included in a large NEH proposal for archival digital audio processing for LAP, and conversely LAP was adopted as the large-scale test for the enhancement of LICHEN at Oulu. An operational version of LICHEN that might be used for LAP was available and demonstrated at DH2007 (Urbana). Further development occurred in conjunction with DH2008 (Oulu), leading to testing of LICHEN on LAP materials in Georgia in Fall 2008. As a result of these steps, the collaborators rewrote the specifications for what the program needed to do in February 2009, as substantial bodies of archivally-processed LAP sound files became available for LICHEN development and testing. The collaborators decided that the key requirements for the specification arose from not only the characteristics of the data (e.g. the structure consisting of interviews, reels and clips with metadata and multimedia on each level) and the desired uses of the framework (e.g. search, browse, and view possible audio selections on a map with GIS), but also from the fact that both tools development and the final stages of data preparation were taking place simultaneously. The data had to be available as flat files usable through any regular file browser, and the sheer scale of the data ruled out the possibility of creating duplicate files inside the program structure as originally designed. The software needed to make use of the existing files and file structure, and so the task of combining the data and the tools became an exercise in conforming tools to data with as little effect on the data itself as possible. To this end, the collaborators developed two methods for bringing in the data and its associated metadata: 1) a general-purpose parser that could traverse file and folder structures and evaluate regular expression patterns to parse metadata from the file and folder names, and 2) a mechanism for re-formatting standard spreadsheet documents into XML documents for use by the tools. The existence of the tools affected the preparation of the data in two ways: 1) some incorrectly named files and folders had to be renamed to conform to the agreed format, and 2) all text files had to be converted from the Microsoft Word format into a plain text format for viewing

from within the developed tools. Both changes were also beneficial to the data collection itself, improving consistency within the data and file support across different platforms. In turn, the developed tools provide access into the database through browsing of the data by natural entities such as interviews, topics and geographic location (as opposed to just files and folders) and queries leveraging the full potential of all the metadata fields.

The DASS product was launched in April 2009 at the SECOL conference (New Orleans), and public distribution began in the summer of 2009 after resolution of the legal issues of the collaboration with university authorities at Georgia and Oulu. Even given the close collaboration between developers, arriving at legal language that suited the university authorities proved to be quite difficult: the Georgia lawyers thought they could be more laissez-faire because the product was unlikely to generate substantial monetary returns, while the Oulu authorities were more focused on retaining the university's rights of ownership. In the end, separate rights language had to be included for materials developed at Georgia and for LICHEN development at Oulu.

As might be expected, integration of a large-scale database with multimedia display functions, while still maintaining the high degree of usability necessary for public access, has turned out to be more difficult than accomplishing any of the separate tasks. And we are not finished. We are currently building a transcription tool to incorporate into the LICHEN framework, so that the public (as well as professional researchers) can contribute transcriptions of audio files. We also want to increase the Geographic Information System (GIS) functionality into the LICHEN toolbox in order to support map-based selection and visualization schemes, surpassing those implemented on the existing Atlas Web site. Finally, we want incorporate the CATMA concordancing tools, currently being built by a collaborating research team at the University of Hamburg. All of these goals, we trust, will make LAP and other data accessible, not just on portable media, but on the web with LICHEN.

References

Kretzschmar, William A. Jr., Anderson, Jean, Beal, Joan, Corrigan, Karen, Opas-Hänninen, Lisa Lena, Plichta, Bartek (2006). 'Collaboration on Corpora for Regional and Social Analysis'. *Journal of English Linguistics*. 34: 172-205.

Kretzschmar, William A. Jr., Childs, Becky, Dollinger, Stefan (2009). 'Creating Public

Corpora: Accessibility, Copyright and Enhancement, and Human Subjects and Metadata'. *Workshop offered at NNAV 38*. Ottawa, October 22-25, 2009.

LAP – *Linguistic Atlas Projects; includes LAGS and DASS*. <http://us.english.uga.edu/>.

LICHEN – *The Linguistic and Cultural Heritage Electronic Network*. <http://www.lichen.oulu.fi/>.

Re-linking a Dictionary Universe or the Meta-dictionary Ten Years Later

Ore, Christian-Emil

c.e.s.ore@edd.uio.no

University of Oslo, Norway

Ore, Espen S.

e.s.ore@iln.uio.no

University of Oslo, Norway

More than ten years ago what was called a meta-dictionary was proposed as a central part of the framework for a dictionary laboratory at the University of Oslo (Ore 2001). The framework has since functioned as a pivot in the combined lexical database, text corpus and manuscript editing system for *Norsk Ordbok* (Norwegian Dictionary). *Norsk Ordbok* is published in twelve volumes (to be completed in 2014) and provides a scholarly and exhaustive account of the vocabulary of Norwegian dialects and the written language Nynorsk, one of the two official written forms of Norwegian.

The architecture of the dictionary framework described in this paper was based upon both explicit and implicit assumptions - and some of the latter were not only not consciously considered in the construction phase, they have also led to features or lacks of features in the system where we now see the need for change. In this paper we look at problems related to links between the meta-dictionary and the sources and show how some of the problems are solved.

1. The meta-dictionary?

In the 1990s a huge amount of lexicographical source material (dictionaries, slip archives and texts) was made electronically available by a national digitization project. By then *Norsk Ordbok* had produced three volumes out of twelve. Being a project started in 1930 the future of the project was highly uncertain. Thus the original motivation behind the meta-dictionary was to create a common web based interface to the background material by inter-linking the material to a common headword list as a meager substitute for the edited dictionary. A similar approach has later been taken by the Dictionary of Old Norse Prose in Denmark (ONP).

Fortunately, the *Norsk Ordbok* project was refunded and revitalized in 2001. It was decided that the new project should be completely digital. As a result a new version of the meta-dictionary was designed.

An entry in the meta-dictionary can be seen as a folder containing (pointers to) possibly commented samples of word usage and word descriptions taken from the linked databases etc. Each entry is labeled by normalized headword(s), word class information and the actual orthographical standard used. The working lexicographers view the meta-dictionary as an easy access to systematized source material. The chief editors use it as tool for headword selection and in dimensioning the printed dictionary. The database used in the Cobuild project for lemma selection from the corpus, is an early example of such a database (Sinclair 1987).

The *Norsk Ordbok* is a historically oriented dictionary covering the period 1600 to the present. The time span and the focus on dialects make the background material heterogeneous. The oldest sources are glossaries compiled in the 17th/18th centuries, mainly the results of work done by vicars collecting information about their parishioners' language on request from the government in Copenhagen. For the description of the word inventory of the current dialects surveys and especially local dictionaries form valuable sources. The meta-dictionary constitutes a bidirectional network. Thus the historical or dialectal dictionary linked to the system can be used as an entry point to the entire set of information.

2. Building a dictionary net

The traditional systematic overviews of the use of words in context have been alphabetically ordered paper slips with each word in a small context and the source information. The slip collections have gradually been replaced by text corpora. In the *Norsk Ordbok* project the slip collections are digitized and linked to the meta-dictionary, and a new annotating tool for singular language observations has been developed. A standard TEI-encoded text corpus spanning the period 1850 to present is gradually constructed. The results from corpus queries can be stored and linked to the meta-dictionary.

The old and the local dictionaries and glossaries constitute an important source for historical and dialectal word usage respectively. Traditionally such dictionaries and glossaries have been transcribed to paper slips and stored in the slip collection. In the new system the dictionaries could have been included in the corpus. This may be done with the newer dialect dictionaries. The old dictionaries are written in Danish or Latin and would have introduced a lot of linguistic noise in the corpus. As these dictionaries are important documents in themselves it was decided to treat them as individual works documenting the language view of their time.

The modern dialect dictionaries are given an XML-encoding according to TEI's printed dictionary format. The 17th /18th centuries' dictionaries are represented by printed, annotated text editions of the original manuscripts. These editions have been transcribed and given a TEI markup reflecting their structure, generally not compatible with TEI's printed dictionary format. Due to their systematic character, <div>-elements can be used to organize the text into chunks describing words and thematic sets of words. The "headwords" are clearly identifiable and are marked as <w>-elements. The loose structure implies that there may be more than one "headword" in each text chunk. The TEI-texts are stored as blobs in a relational database. The TEI-texts are chopped up according to the entries (dictionaries) and the text chunks (glossaries) and stored together with the headwords (slightly normalized) in a separate table structure.

In the early version of the system, the linking between these sources and the meta-dictionary were on the <entry> level for the local and on the <div> level for the old dictionaries. There was no information about the keyword in the selected dictionary that was used to create a link. In some cases when a <w>-element was removed an invalid link from the meta-dictionary to the external text sets was left. Today the link is annotated with the actual headword and the person responsible for the link.

The process is automatic with a manual check: a daily job runs through registered dictionaries and looks for keywords in a special metadata field in the database. If the word is found but there is no existing link between the meta-dictionary and this text unit, a link is created, and the record in the meta-dictionary is marked as changed and will be forwarded to an editor for approval. If the word is not found in the meta-dictionary, a new entry is created and linked with the text unit, and this will be sent to the editor for approval (see also Fournier 2001 and Gärner 2008 for interlinking of dictionaries).

3. The meta-dictionary and other dictionary nets

What constitutes a word is an unresolved linguistic question. A traditional monolingual dictionary is a word form oriented index to a set of concepts and meetings where each entry is indexed by a headword and contains a possible meaning hierarchy with samples. Word forms denoting related concepts are connected by cross references. The Wordnet approach is to focus on the concepts and collect the word forms denoting the same concepts in sets of synonyms (synsets). The synsets can then be organized according to a predefined ontology such as in the Global Wordnet Grid (Vossen 2009).

The two ways of organizing word information is fully compatible. A word net can be converted to a traditional dictionary and a well organized dictionary rich on semantic references can be converted into a Wordnet.

The current meta-dictionary was pragmatically designed 8 year ago. It has in itself become a valuable lexicographical documentation system. The source material spans both in time and space. Due to the practical purpose, that is, editing a traditional dictionary, the word forms are linked mostly etymologically. Thus an entry covers many concepts as does an entry in a traditional dictionary. However, the meta-dictionary also groups how different scholars have described the meaning of a word from 1600 to the present. The resources comprise the old digitized paper slip collections, the dictionaries and glossaries and stored results from querying the corpus. They all represent collections of systematized language documentation in their own right and premises. The entries in the *Norsk Ordbok* are in fact just yet another source of (scientifically) systematized language information linked to the others. However, the *Norsk Ordbok* system groups the information according to meaning and the dictionary is rich in synonym relations. A future research project is to use the information from the dictionary to create a second set of articles in the meta-dictionary in a Wordnet fashion with semantic relations.

TEI, Text Encoding Initiative. <http://www.tei-c.org/P5/>.

Norsk Ordbok. <http://www.no2014.uio.no>.

ONP, Dictionary of Old Norse Prose. <http://www.onp.hum.ku.dk>.

WordNet. <http://wordnet.princeton.edu/>.

References

Fournier, Johannes (2001). 'New Directions in Middle High German Lexicography: Dictionaries Interlinked Electronically'. *Literary and Linguistic Computing*. **16/1**: 99-111.

Gärtner, Kurt (2008). 'The New Middle High German Dictionary and its Predecessors as an Interlinked Compound of Lexicographical Resources'. *DH2008 conference*. Oulu, Finland, 2008.

Ore, Christian-Emil (2001). *Metaordboken - et elektronisk rammeverk for Norsk Ordbok?* Gellerstam, Martin et al. (ed.). Nordiska studier i leksikografi. Göteborg. V. 5.

Sinclair, J. M. (ed.) (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

Vossen, Piek (2009). 'From Wordnet, EuroWordNet to the Global Wordnet Grid'. *eLEX2009 conference*. Louvain-la-Neuve, Belgium, October 22-24.

Digital Resources for Art-Historical Research: Critical Approach

Rodríguez Ortega, Nuria

nro@uma.es

Universidad de Málaga (Spain)

Esta comunicación parte de una convicción: la necesidad de empezar a construir un discurso crítico y autoconsciente que reflexione sobre la incidencia que el medio digital, la tecnología informática y los entornos web interactivos están teniendo en los procesos de investigación del hecho artístico.

Después de casi dos décadas utilizando Internet y la Web, cualquier cosa que se pueda decir sobre cómo el medio digital ha facilitado y optimizado estos procesos de investigación y la adquisición de información puede sonar a obviedad. Efectivamente: el acceso a la información a través de recursos digitales es rápida; muy rápida. Es ubicua: podemos conseguir información desde cualquier lugar y en cualquier momento. Es global: podemos buscar información en cualquier parte del mundo, y encontrarla - siempre y cuando haya sido previamente digitalizada -. Es más flexible: el investigador-usuario puede seleccionar qué tipo de información obtener, cuándo y de qué modo. Y lo que constituye uno de los aspectos más interesantes: la información recuperada en el medio digital es visualmente más significativa debido a la versatilidad de la interfaz gráfica, el soporte multimedia y el lenguaje hipertextual. Pensemos en un ejemplo tan simple como la posibilidad de visualizar paralelamente dos textos o dos imágenes.

La investigación a través del medio digital es más fácil, más significante y proporciona más posibilidades al especialista para profundizar en sus investigaciones. Asimismo, el uso de recursos digitales transforma las prácticas investigadoras y condiciona determinadas modalidades de estudio. Sin embargo, en el campo de la Historia del Arte todo esto lo sabemos de manera inmediata e intuitiva, pues todavía no hemos empezado a desarrollar un discurso crítico que reflexione sobre el modo como lo digital está incidiendo en nuestra área de conocimiento.

Hay que tener en cuenta que el uso de recursos digitales no es neutro, sino que conlleva - implícita o explícitamente, consciente o inconscientemente - determinadas conceptualizaciones sobre el objeto de estudio, y propende a determinados modos de aproximación y análisis. De hecho, el uso de

recursos digitales se mueve en una especie de ambivalencia entre la utilidad incuestionable y las recategorizaciones implícitas. No olvidemos que toda página web o plataforma digital es un discurso, y como tal implica una determinada conceptualización del hecho, fenómeno o idea sobre el que se construye dicho discurso.

Un ejemplo simple y clarificador nos lo proporciona el empleo de imágenes en formato gigapíxeles, que se han popularizado sobre todo a partir de diversas iniciativas llevadas a cabo entre instituciones culturales y Google Earth. La posibilidad de visualizar con detalle nunca antes imaginable determinadas secciones de una imagen pictórica constituye una inestimable ayuda para analizar con rigurosidad elementos y componentes que hasta ahora habían sido marginales. Asimismo, la posibilidad de seccionar una imagen para analizar una o varias de sus partes favorece con mucho la tarea del historiador del arte, buena parte de la cual consiste en examinar imágenes artísticas. Ahora bien, esta posibilidad de análisis también está favoreciendo una aproximación a la imagen artística basada en el fragmento y en el detalle, mientras que la aprehensión de la imagen como totalidad unitaria tiende a perderse. Hay que tener en cuenta que en el medio digital, la imagen como entidad autónoma no existe de la misma manera que existe en el medio analógico. En el medio digital, lo que tenemos es una masa de píxeles que nosotros podemos manipular a nuestro antojo: cortando, disgregando, remezclando, interviniendo..., o en función de las posibilidades que nos permite el software y la capacidad de la interficie gráfica. Además, el valor de cualificación de una imagen digital reside, precisamente, en su capacidad de disección, intervención y ampliación. Eso es lo que busca un usuario que trabaja con imágenes digitales.

En un escrito de 2005, W. Vaughan indicaba que la reproducción completa de un objeto en la pantalla del ordenador únicamente esquema nemotécnico, una especie de guía para el espectador en su búsqueda del detalle (Vaughan, 2005: 6). Si observamos, por ejemplo, el modo como Google Earth Prado nos muestra sus imágenes en alta resolución, parece que Vaughan da en el clavo en sus apreciaciones [v. fig. 1].



Fig.1: Imagen de Google Earth Prado

Podríamos pensar que el historiador del arte dispone de los elementos críticos suficientes para posicionarse ante estas imágenes como lo que son: instrumentos llevar a cabo la investigación histórico-artística. Sin embargo, ¿qué sucede con el público no experto? Si atendemos a una de las funcionalidades que ha incorporado Flickr: la posibilidad de fragmentar las fotografías y comentarlas trozo a trozo, parece que estamos ante la constatación de que el medio digital está promocionando un acercamiento fragmentario a la imagen artística [fig. 2].



Fig. 2: Captura de pantalla de Flickr

Otro ejemplo interesante lo constituyen los catálogos y las bases de datos, uno de los recursos digitales más utilizados en el campo de la Historia del Arte como instrumento de información, y uno de los más desarrollados. Sin poner en cuestión la utilidad de estos bancos y catálogos, no debemos perder de vista que la imagen que proyectan es la de una sucesión de registros en los que el objeto artístico «aparece» solo, por más que aquél incluya algún campo para la descripción crítica del objeto. Es lógico que las bases de datos funcionen así, puesto que uno de sus objetivos es organizar las informaciones en compartimentos estructurados. Sin embargo, esta imagen de objetos aislados unos detrás de otros no se corresponde en absoluto con la aproximación contextualista y culturalizadora que desde hace décadas prima como modelo interpretativo de los

hechos artísticos. Esto no quiere decir que las bases de datos y los catálogos sean inoperantes; todo lo contrario, constituyen eficientes herramientas de información. Pero si el medio digital está copado mayoritariamente por estas herramientas, y la cultura tecnológica del historiador del arte - muchas veces precaria o inexistente - no le impulsa a indagar en otro tipo de recursos, la visión de los hechos artísticos que estaremos legando al futuro será fundamentalmente una visión aislacionista y descontextualizada.

La ausencia de una reflexión consciente sobre las especificidades del medio digital también ha supuesto que, hasta la fecha, estemos infráutilizando, por no decir mal-utilizando, los recursos digitales. Muchas de estas potencialidades están todavía sin explorar. En numerosas ocasiones, lo que encontramos son transposiciones electrónicas de nuestro mundo analógico. Así, no es difícil encontrar páginas web en las que se contienen enciclopédicos con los mismos mecanismos de lectura - lineal y discursiva - que en una edición impresa, que el modo de lectura en Internet está basado en el relato hipertextual y en el lenguaje hipermedia. [fig. 3].



Fig. 3: Catálogo online del Museo del Prado con contenido lineal y discursivo

Un caso interesante al respecto es el Museo Virtual de Arte Uruguay,¹ un museo enteramente virtual que ha recibido el beneplácito de la comunidad internacional por su carácter innovador. Sin embargo, este museo sigue siendo un museo, y funciona como tal: exposiciones, galerías, etc. El «museo», no obstante, es una categoría mental y cultural que forma parte de nuestro universo analógico. En el medio digital, abierto, interactivo, en red, con múltiples posibilidades de *display* multimedia y de elaboración de relatos curatoriales multisecuenciales, la categoría «museo», como la conocemos, no tiene sentido. Podemos deducir que se ha escogido esta categoría como marco contextualizador porque el museo es una noción que le da al objeto «musealizado» un marchamo y una aceptación social de hecho

artístico (Vaughan, 2005: 8); y también porque es una categoría mental con la que el público está familiarizada, que le ayuda a «entender» perfectamente esta propuesta virtual. En este caso, pues, el medio digital se adapta a prácticas ya existentes y tradicionales en el mundo no digital. Por una parte constituye una solución efectiva, porque hace más comprensible la experiencia, pero con ella también pierde la oportunidad de explorar las características del medio digital para proponer nuevas modalidades de exhibición artística y de diseño de narrativas.

De hecho, algunos de los recursos aplicados en el campo de la Historia del Arte, como las reconstrucciones 3D o las espacializaciones virtuales, que hemos celebrado durante años como grandes avances de la tecnología, están contribuyendo a extender la idea de que el medio digital es la transposición electrónica del mundo analógico, obviándose que el medio digital tiene sus propia lógica de funcionamiento y se rige por principios diferentes a los de la realidad física y material.

La presentación expondrá con más detalle otros ejemplos que nos invitan a tomar conciencia de la necesidad de desarrollar una reflexión crítica sobre la presencia de la Historia del Arte en el medio digital y, sobre todo, a repensar qué tipo de Historia del Arte “digital” estamos construyendo y legando al futuro. Una reflexión que tiene un corolario último y que evoca algunas consideraciones recientes de J. Drucker (2009) y C. Borgman (2009): la inevitable implicación del historiador del arte en los procesos de diseño, desarrollo e implementación de los recursos digitales con los que construir el conocimiento histórico-artístico.

References

- Borgman, C. L.** (2009). *Scholarship in the Digital Age: Blurring the Boundaries between the Sciences and the Humanities*. College Park, MD: Maryland Institute for Technology in the Humanities. <http://works.bepress.com/borgman/216/> (accessed 27 February 2010).
- Brea, J. L.** (2008). *Cultura_Ram. Mutaciones de la cultura en la era de su distribución electrónica*. Barcelona: Gedisa.
- Campàs, J.** (2006). 'Estudio hipertextual de Las Meninas de Velázquez'. *1st SEEArcWeb Workshop on Open and Distance Learning (ODL) Strategies*. Barcelona: Athens, pp. 85-114.
- Drucker, J.** (2009). 'Blind Spots. Humanist must plan their digital future, The Chronicle Review'. *The Chronicle of Higher Education*. 55 (30). <http://chronicle.com/free/v55/i30/30b00601.htm> (accessed 27 February 2010).

Frischer, B. (2009). 'Art and Science in the Age of Digital Reproduction'. *Mimetic Representation to Interactive Virtual Reality*. Sevilla, 17-20 de junio.

Rodríguez Ortega, N. dir. (2009). *Teoría y literatura artística en la sociedad digital. Diseño y aplicabilidad de colecciones textuales informatizadas*. Gijón: Trea.

Vaughan, W. (2005). 'History of Art in the Digital Age. Problems and Possibilities'. *Digital Art History. A Subject in Transition. Exploring Practice in a Network Society*. Bentkowska-Kafel, A., Cashen, T., Gardiner, H. (eds.). UK: Intellect Books, pp. 3-13.

Notes

1. <http://muva.elpais.com.uy/>

Towards Hermeneutic Markup: An architectural outline

Piez, Wendell

wapiez@mulberrytech.com

Mulberry Technologies, Inc., USA

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

Hermeneutic markup in its full form does not yet exist. XML, and especially TEI XML, provides a foundation for this work. But due to limitations both in currently dominant approaches to XML, and in XML itself, a number of important desiderata remain before truly sophisticated means can be made available for scholars to exploit the full potentials of markup for literary study, as implied, for example, by ideas such as Steven Ramsay's Algorithmic Criticism or what I described in 2001 (following Rockwell and Bradley) as "*exploratory markup*" (Piez 2001. See also especially Buzzetti, 2002 and McGann, 2004).

Prototype user interfaces designed to enable one or another kind of *ad hoc* textual annotation or markup have been developed, for the most part independently of one another (several are cited.). This shows that the idea of hermeneutic markup, or something like it, is not new; but none of these have yet made the breakthrough. An important reason is that hermeneutic markup in its full sense will not be possible on the basis simply of a standard tag set or capable user interface, because it will mean not just that we can describe a data set using markup (we can already do that), but that we can actively develop, for a *particular* text or family of texts, an

appropriate, and possibly highly customized, means and methodology for doing so.

A demonstration of a prototype markup application helps to show the potentials and challenges, in a very rudimentary form [screenshots appear in Figure 1.] This graphical and interactive rendering of markup in the source files presents an interpretation of the grammatical/rhetorical structure (sentences and phrases) as well as verse structure (lines and stanzas) in the text. Unfortunately, while the encoding for the sonnets here is not inordinately difficult – "milestones" are used, in a conventional manner, to denote the presence of structures that overlap the primary structure of the encoded document – the code that renders it (not included in the package) incurs significant extra overhead to run, because XML technologies are ill-fitted to manage the kind of information we are interested in here, namely the overlapping of these structures that characterizes the sonnet form. XML doesn't do overlap. As long as a sentence or phrase overlaps a line – a very common occurrence and important poetic device – the normative XML data model, a "tree", cannot capture both together. In order to do processing like what happens here, one or another workaround has to be resorted to. So while XML is being used here, it is a clumsy means to this end.

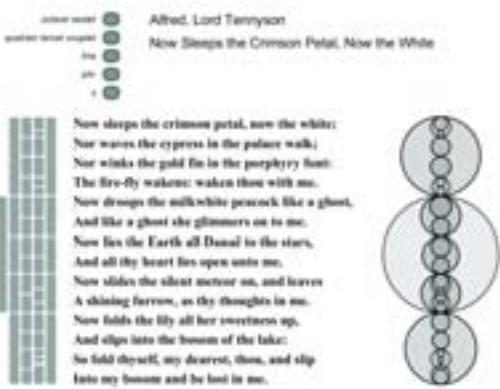


Figure 1: Screenshots of three sonnets with rendition of overlapping (verse and sentence/phrase) structures. The interface (implemented in W3C-standard SVG) is dynamic and responds to user input to highlight overlapping ranges of text.

But overlap is only part of the problem. Consider Alfred Lord Tennyson's *Now Sleeps the Crimson Petal, Now the White*. This too is a sonnet, after a fashion, although it does not have a conventional sonnet's octave/sestet structure. Since this application does not work with a schema, this is not a problem here. Yet as texts or collections grow in scale and complexity, having a schema is essential to enforcing consistency and helping to ensure that like things are marked up alike. A framework for this application must not only find a way to work around the overlap; it must also deploy a schema (or at any rate some sort of validation technology) flexible enough – at least if this instance is to be valid to it – that such outliers from regular form are permissible, even while attention is drawn to them (see Birnbaum 1997).

Currently, XML developers generally (setting aside the problem of overlap) do not consider this to be problematic in itself; indeed, part of the fun and interest of markup is in devising and applying a schema that fits the data, however strange and interesting it may be. What is not so fun is to have to repeat this process endlessly, being caught in a cycle of amending and adjusting a schema constantly (and sooner or later, scripts and stylesheets) in order to account for newly discovered anomalies. Sooner or later, when exhaustion sets in or the limits of technical knowhow are reached, one ends up either faking it with tags meant for other purposes (thereby diluting markup semantics in order to *pretend* to represent the data), or just giving up.

Extending a schema is found to be a problem not only because validating and processing any model more complex than a single hierarchy is a headache even for technical experts, but also, more generally, because current practices assume a top-down schema development process. Despite XML's support for processing markup even without a

schema, both XML tools and dominant development methodologies assume that schema design and development occurs prior to the markup and processing of actual texts. This priority is both temporal and logical, reflecting a conception of the schema as a declaration of constraints over a set of instances (a “type”), appropriate to publishing systems built to work with hundreds or thousands of documents, with a requirement for backwards compatibility (documents encoded earlier cannot be revised easily or at all) and limited flexibility to adapt to new and interesting discoveries. The centrality of the schema within this kind of system inhibits, when it does not altogether frustrate, the flexible development of a markup practice that is sensitive, primarily, to a text under study, and this conception of a schema's authority works poorly when considering a single text *sui generis* – the starting point for hermeneutic markup. In hermeneutic markup, a schema should be, first and last, an apparatus and a support, not a Procrustean bed.

All these problems together indicate the outline of a general solution:

- A data model supporting arbitrary overlap.
- Interfaces, including a markup syntax, that facilitate the creation, editing and analysis of texts using this data model, with the capability of defining *ad hoc* elements and properties (attributes) on the fly.
- A transformation technology supporting (in addition to data transformations) analytical tools applicable to the markup as such (not just the raw text), with the capability of managing elements and their properties in sets, locating them, listing them by type, sorting, visualizing and comparing them.
- Schema-inferencing capabilities for describing the structural relations within either an entire marked-up corpus, or within identifiable segments, sections or profiles of it.
- In connection this, a schema technology that supports partial and modular validation.

A system with all these features would support an iterative and “*agile*” approach to markup development. We would start by tagging. (In a radical version of this approach we might start by tagging for presentation, perhaps using just a lightweight HTML or TEI variant for our first cut.) Then we introduce a provisional schema or schemas capable of validating the tagging we have used. This requires assessing which cases of overlap in the text are essential to our document analysis, and which are incidental and subject to normalization within hierarchies. Having refined the schema, we return to the tagged text, to consider both how its tagging falls short (with

respect to whatever requirements we have for either data description or processing), and how it may be enhanced, better structured and regularized. During this process we also begin to develop and deploy applications of the markup. We then revise, refactor and extend both tagging and schema, applying data transformations as needed, in order to better address the triple goals of adequate description, processing, and economy of design.

Such a system would not only be an interesting and potentially ground-breaking approach to collaborative literary study; it would also be a platform for learning about markup technologies, an increasingly important topic in itself. Moreover, hermeneutic markup represents an opportunity to capitalize on investments already made, as texts encoded in well-understood formats like TEI are readily adaptable for this kind of work.

Many of these capabilities have already been demonstrated or sketched in different applications or proposals for applications, including W3C XML Schema (partial validation); James Clark's Trang (schema inferencing for XML); LMNL/CREOLE (overlap, structured annotations, validation of overlap); JITTs (XML "profiles" of concurrent overlapping structures); and TexMECS (overlap, "virtual" and discontinuous elements).

The presentation will conclude with a demonstration of various outputs from the data sources used in the demo, which provide building blocks towards the kind of system sketched here. A range-analysis transformation can show which types of structures in a markup instance overlap with other structures, and conversely which structures nest cleanly. Complementary to this, an "*XML induction processor*" is capable of deriving well-formed XML representations of texts marked up with overlapping structures – from which, in turn, XML schemas can be derived.

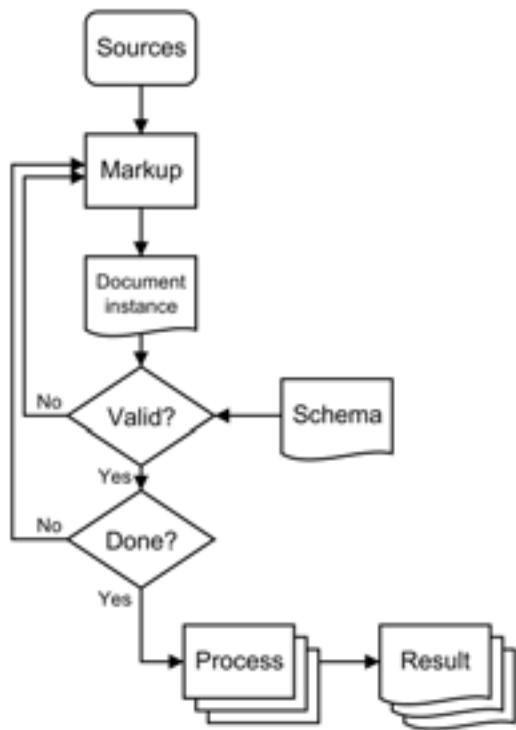


Figure 2: A workflow diagram showing the architecture of present (XML-based) markup systems. Both schema and processing logic are considered to be static; modifying them is an activity extraneous to document markup and production.

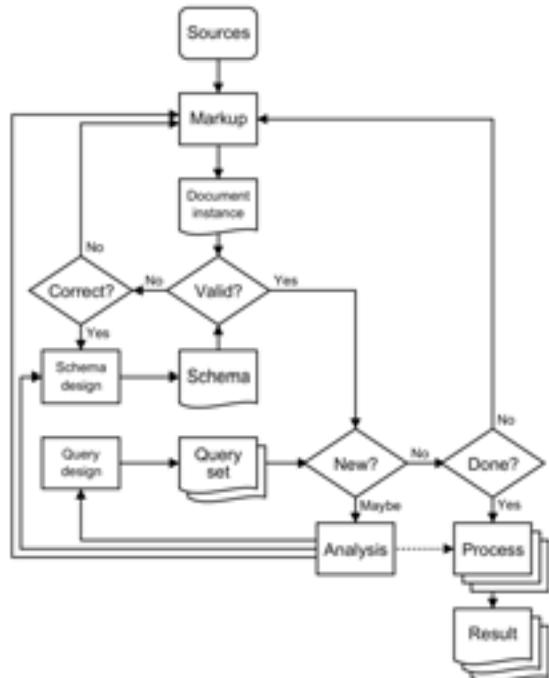


Figure 3: An architecture capable of supporting hermeneutic markup would account directly for document analysis and for the design of schema, queries and processing. While in fact this

is often done even today, one has to work against the current tool set to do it, questioning its assumptions regarding the purposes, roles and relations of source text, markup and schema. A final version of this paper, with the demonstrations, is available at <http://piez.org/wendell/papers/dh2010/index.html>

References

- Birnbaum, David.** 'In Defense of Invalid SGML'. *ACH/ALLC*. Kingston, Ontario, 1997.
- Buzzetti, Dino** (2002). 'Digital Representation and the Text Model'. *New Literary History*. **33(1)**: 61-88.
- CATMA**. *University of Hamburg (Jan Christoph Meister)*. <http://www.jcmeister.de/html/catma-e.html>.
- Caton, Paul.** 'LMNL Matters? (presenting the Limner prototype tagging platform)'. *Extreme Markup Languages 2005*. Montréal, Québec, 2005.
- Czmiel, Alexander.** 'XfOS (XML for Overlapping Structures)'. *ACH/ALLC 2004*. Göteborg, Sweden, 2004.
- Di Iorio, Angelo, Silvio Peroni and Fabio Vitali.** 'Towards markup support for full GODDAGs and beyond: the EARMARK approach'. *Balisage: The Markup Conference 2009*. Montréal, Canada, 2009. <http://www.balisage.net/Proceedings/vol3/html/Peroni01/BalisageVol3-Peroni01.html>.
- Durusau, Patrick, and Matthew Brooke O'Donnell.** 'Coming down from the trees: Next step in the evolution of markup? (Presenting JITTs, Just-in-time Trees)'. *Extreme Markup Languages 2002*. Montréal, Québec, 2002. http://www.durusau.net/publications/NY_xml_sig.pdf.
- Huitfeldt, Claus.** *Markup Languages for Complex Documents (MLCD)*. <http://decentius.aksis.uib.no/mlcd/en.htm>.
- Image Markup Tool. University of Victoria (Martin Holmes).** http://tapor.uvic.ca/~mholmes/image_markup/.
- Lancashire, Ian** (1995). *Early Books, RET Encoding Guidelines, and the Trouble with SGML*. <http://www.ucalgary.ca/~scriptor/papers/lanc.html>.
- LMNL wiki**. http://www.lmnl.org/wiki/index.php/Main_Page.
- McGann, Jerome** (2004). 'Marking Texts of Many Dimensions'. *A Companion to Digital Humanities*. Susan Schreibman, Ray Siemens, John Unsworth (ed.). Oxford: Blackwell.
- NINES project. University of Virginia (Jerome McGann, et al.)**. <http://www.nines.org>.
- Piez, Wendell** (2001). 'Beyond the Descriptive vs Procedural Distinction'. *Markup Languages: Theory and Practice*. **3(2)**. <http://www.piez.org/wendell/papers/beyonddistinction.pdf>.
- Ramsay, Steven** (2008). 'Algorithmic Criticism'. *A Companion to Digital Literary Studies*. Susan Schreibman and Ray Siemens (ed.). Oxford: Blackwell. <http://www.digitalhumanities.org/companionDLS/>.
- Sperberg-McQueen, C. Michael** (1991). 'Text in the electronic age: Textual study and text encoding, with examples from medieval texts'. *Literary & linguistic computing*. **6(1)**: 34-46.

Works, Documents, Texts and Related Resources for Everyone

Robinson, Peter

p.m.robinson@bham.ac.uk

Institute for Textual Scholarship, University of Birmingham

Meschini, Federico

fmeschini@dmu.ac.uk

Centre for Textual Studies, De Montfort University

A common trope in discussions of scholarly editions in digital form is to praise, on the one hand, the extraordinary potential of electronic editions while, on the other hand, regretting that so few actual electronic editions come anywhere near realizing this potential (Robinson 2005). The potential is well-known: an explicit hyper-textual structure, publication in a distributed network environment, escape from the storage limit of the printed medium and possession of multiple layout possibilities (such as normalized and diplomatic transcriptions juxtaposed to facsimile images).

The difficulties are also well-known: among them, the need for a formal, comprehensive and efficient encoding scheme to underpin scholarly editions in electronic form. The Text Encoding Initiative Guidelines provided a crucial element, by supplying namings, specifications and structure for key components of electronic editions: thus the specialized lower-level elements for manuscript description and critical apparatus, along with higher-level elements such as msDescription and facsimile. However, the TEI does not address two areas, crucial for the full encoding of scholarly editions in electronic form:

1. The naming of components of the editions: thus, of the works edited and their parts; the source manuscripts or print documents and their parts which carry the texts of the work edited;
2. The relationships between the components: thus, between the documents, the texts they carry, and the works which those texts instance.

This paper reports on a scheme prepared by the authors, designed to provide a solution to the problems proposed in both areas. The provision of a shared epistemological framework for handling works, texts and text sources (cf. Buzetti 2009) will also facilitate the shift from stand alone publishing frameworks to shared distributed online environments, enabled by powerful and flexible

underlying infrastructures,¹ generally named Virtual Research Environments (Fraser 2005, Dunn et al. 2008).

This framework will advance interoperability, long a problem area in electronic texts. Interoperability has been defined by IEEE as “The ability of two or more systems or components to exchange information and to use the information that has been exchanged”.² A recent briefing paper by Gradmann identifies four different levels of interoperability, one built on the top of the other. From the bottom these levels are technical/basic, syntactic, functional and semantic. While technologies such as TCP/IP, HTTP and XML already provide sound basis for interoperability at the lower levels, much work is still to be done at the top levels. The semantic frame for interoperability offered by this scheme speaks to this need.

Semantic issues in networked publication systems are advanced by the work done in the last years on the ‘Semantic Web’ (Berners-Lee et al. 2001), which has recently evolved into the Linked Data initiative (Berners-Lee 2006). The Semantic Web seems to have survived its own hype, having finally entered the plateau of productivity phase, as happened for XML some years ago. The ontological level of the Semantic Web stack, represented by the OWL language, has presented a steep learning curve, due partly to its roots in Description Logic and First-Order Logic (Gruber 1993), but also presents at the same time the greatest potential.

The relationship between textual scholarship in its electronic dimension and ontologies has not hitherto been much apparent, as textual scholars using digital methods have focussed rather on the related, but separated field of Library and Information Science (Vickery 1997). However, ontologies have much to offer the textual editing enterprise. Both ‘recensio’ and the construction of a stemmatic graph are implicit formalizations that would benefit from the adoption of an explicit modelling. Moreover, both Sperberg-McQueen and Peter Shillingsburg implicitly hints at the potentialities of an ontological approach in scholarly editions, the former when writing about the “infinite set of facts related to the work being edited” (Sperberg-McQueen 2002) and the latter about “electronic knowledge sites” (Shillingsburg 2006).

In the world of digital humanities and electronic editions proficient uses of ontologies have already appeared, such as the Discovery³ and the Nines⁴ projects, also leveraging existing standards from related sectors such as IFLA’s FRBR⁵ (Mimno 2005) or the cultural heritage oriented CIDOC-CRM⁶ (Ore et al. 2009).

Substantial work is now being done on implementing an actual interchange and interoperability framework for electronic editions, and arbitrary portions of them, of the kind, in (for example) the COST Action Interedition.⁷ A first proposal by Peter Robinson (Robinson 2009) was based on the Kahn/Wilensky Architecture (Kahn et al. 1995),⁸ having therefore a naming authority together with a series of key/value pairs identifying portions of an electronic text, which therefore could be exchanged over the net thanks to a protocol such as the one established by the OAI-PMH standard.⁹ This addressed the first need stated above, for agreed conventions on naming. The second need, for formal expression of relationships, is addressed by the adoption of the Linked Data paradigm. While keeping the use of the Kahn/Wilensky Architecture for the labelling system, and using a URN-like syntax compatible with the Semantic Web requirements, an ontology representing the entities involved together with their relationships has been developed.

The main entities of this ontology are:

- 'Work': Canterbury Tales, and 'WorkPart', the first line of the Canterbury Tales;
- 'Document', the Hengwrt or the Ellesmere manuscripts, and 'DocumentPart', a page, folio or quire, which might carry an instance of the 'Work'
- 'Text': a single instance of a work, or work part, in a document or document part. Thus: the *text* of the *work* 'The Canterbury Tales' as it appears in the *document*, the Hengwrt manuscript;

The three-fold distinction between 'Work', 'Document' and 'Text' reflects the fundamental scholarly distinction between the 'Work', independent of its realization in any object; the 'Document' which might carry an instance of the 'Work'; and the 'Text': the instance of the work in the document. Digital resources such as 'Image' or 'Transcript' are related to 'Text' and 'Document' and their parts, using relationship such a 'hasImage', 'isTranscriptOf', or 'transcribedFrom'. Basic properties such as "isPartOf" or other properties from existing vocabularies, such as Dublin Core,¹⁰ have also been used, so to guarantee compatibility with other schemes in the best possible way. The resulting RDF can be stored in a triplestore and made available on the web, so to allow further uses from third parties without the need to establish exclusive protocol verbs.

This paper will present the methodological thinking behind the development of this ontology for the interchange of electronic editions of literary texts, starting from the first proposal until the more recent developments. The ontology will be contextualized with the existing related standards, particularly

FRBR, CIDOC-CRM and the recent OAI-ORE¹¹ (a gross-grained vocabulary for the reuse and exchange of digital objects developed by the Open Access Initiative) and with the similar initiative of the Canonical Text Service Protocol (CTS),¹² which recently also added an ontological dimension to its basic syntax (Romanello et al. 2009).

References

- Berners-Lee, T., Hendler, J., Lassila, O.** (2001). 'The Semantic Web'. *The Scientific American*. **May 2001**: 34–43.
- Berners-Lee, T.** (July 2006). *Linked Data*. <http://www.w3.org/DesignIssues/LinkedData.html>.
- Buzzetti, D.** (2009). 'Digital Editions and Text Processing'. *Text Editing, Print and the Digital World*. Deegan, M., Sutherland, K. (eds.). Ashgate, pp. 45–61.
- Dunn, S., Blanke, T.** (2008). 'Next Steps for E-Science, the Textual Humanities and Vres'. *D-Lib Magazine*. **1/2**. <http://www.dlib.org/dlib/january08/dunn/01dunn.html>.
- Fraser, M.** (2005). 'Virtual Research Environments: Overview and Activity'. *Ariadne*. **44**. <http://www.ariadne.ac.uk/issue44/fraser/>.
- Gradmann, S..** *INTEROPERABILITY. A key concept for large scale, persistent digital libraries*. <http://www.digitalpreservationeurope.eu/publications/briefs/interoperability.pdf>.
- Gruber, T.R.** (1993). 'A translation approach to portable ontology specifications'. *Knowledge Acquisition*. **5**: 199–220.
- Kahn, R., Wilensky R.** (May 1995). *A Framework for Distributed Digital Object Services*. <http://www.CNRI.Reston.VA.US/home/cstr/arch/k-w.html>.
- Mimno, D., Crane G., Jones, A.** (2005). 'Hierarchical Catalog Records: Implementing a Frbr Catalog'. *D-Lib Magazine*. **10**. <http://www.dlib.org/dlib/october05/crane/10crane.html>.
- Ore, C., Eide, Ø.** (2009). 'TEI and cultural heritage ontologies: Exchange of information?'. *Literary and Linguistic Computing*. **2**: 161–172. <http://llc.oxfordjournals.org/cgi/content/abstract/24/2/161>.
- Robinson, P.** (2005). 'Current Issues in Making Digital Editions of Medieval Texts—or, Do Electronic Scholarly Editions Have a Future?'. *Digital Medievalist*. <http://www.digitalmedievalist.org/journal/1.1/robinson/>.

Robinson, P. (2009). 'Electronic Editions for Everyone'. *Text and Genre in Reconstruction*. McCarty. W. (ed.). Cambridge: Open Book Publishing, pp. 183-201.

Romanello, M., Berti, M., Boschetti, F., Babeu A., Crane G. (2009). *Rethinking Critical Editions of Fragmentary Texts by Ontologies*. Milano, Italy: Elpub. <http://conferences.aepic.it/index.php/elpub/elpub2009/paper/view/158>.

Sperberg-McQueen, C. M. (2002). *How to Teach Your Edition How to Swim*. <http://www.w3.org/People/cmsmcq/2002/cep97/swimming.xml>.

Shillingsburg, P. (2006). *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge: Cambridge University Press.

Vickery, B. C. (1997). 'Ontologies'. *Journal of Information Science*. 4: 277-286. <http://jis.sagepub.com/cgi/content/abstract/23/4/277>.

Notes

1. Such as the European initiative DARIAH <<http://www.dariah.eu>>
2. <<http://en.wikipedia.org/wiki/Interoperability>>
3. <<http://www.discovery-project.eu>>
4. <http://www.nines.org>
5. <http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records>
6. <http://cidoc.ics.forth.gr>
7. <http://www.interedition.eu>
8. Which constitutes also the basis for the Handle system <http://www.handle.net>
9. <http://www.openarchives.org/OAI/openarchivesprotocol.html>
10. <http://dublincore.org>
11. <http://www.openarchives.org/ore>
12. <http://chs75.chs.harvard.edu/projects/diginc/techpub/cts>

A Day in the Life of Digital Humanities

Rockwell, Geoffrey

geoffrey.rockwell@ualberta.ca

Philosophy and Humanities Computing, University of Alberta, Canada

Ruecker, Stan

sruecker@ualberta.ca

English and Humanities Computing, University of Alberta, Canada

Organisciak, Peter

organisc@ualberta.ca

Humanities Computing, University of Alberta, Canada

Meredith-Lobay, Megan

megan.meredith-lobay@ualberta.ca

Arts Resource Centre, University of Alberta, Canada

Ranaweera Kamal

kamal.ranaweera@ualberta.ca

Arts Resource Centre, University of Alberta, Canada

Sinclair, Stéfan

sgs@mcmaster.ca

McMaster University, Canada

On March 18th, 2009 over 90 people participated in a collaborative documentation project called *A Day in the Life of Digital Humanities*. The participants blogged what they did that day in the spirit of digital humanities as a form of autoethnography that could help answer the question, "just what do we do?"

In this paper we will:

- Discuss the conception, design and delivery of the project,
- Discuss the intellectual property paradigm that we adopted to make this project one that produces open documentation for use by other projects,
- Reflect on the lessons learned about such social research projects and theorize the exercise, and
- Discuss the second Day of DH project that, building on the strengths of the first, will be run March 18th, 2010.

1. From Conception to Delivery

The original idea for the project was to develop a communal response to questions asking exactly what it is that we do in the digital humanities. In 2006, "The State of Science & Technology in

Canada" from the Council of Canadian Academies reported humanities computing as an emerging field of strength in Canada. Since then, there have been requests in various forms for an explanation of what the previously unnoticed field was.¹

The form of the response was inspired by a lecture by Edward L. Ayers (currently now President of the University of Richmond) that we had heard about, titled "What Does a Professor Do All Day, Anyway?" Ayers was an early computing historian whose "The Valley of the Shadow" project was one of the two founding IATH projects. In that lecture, he reflected on how people, including his own young son, know little about what a professor does. As he put it,

"In the eyes of most folks, a professor either portentously and pompously lectures people from his narrow shaft of specialized knowledge, or is a bookworm – nose stuck in a dusty volume, oblivious to the world."²

The situation is even worse in the digital humanities, where not only do people not know what we do as academics, they also don't know what "humanities computing" or the "digital humanities" are. It's not even clear if practitioners agree with each other on these basic matters. Ayers's approach to answering this question was the simplest and most cohesive: simply to describe each part of his day, task by task. A Day in the Life of Digital Humanities scales this approach up to a participatory project. We hoped to address the questions about the nature of digital humanities academic work by reflecting as a community.

The Day of DH (as we call it) was thus conceived to provide one form of response to the definition of the field: not through speculation, but through observation. In this context we will also briefly demonstrate the WordPress setup and the wiki that was used to coordinate materials.³

2. Intellectual Property Paradigm: Collaborative Publishing

As for all projects with human participants in Canadian academia, we first had to apply for ethics review. We presented the project not simply as a study of what the participants are doing, but as a collaborative publication. The paradigm therefore was that we were organizing a collective documentation project where the results would be a form of publication that would be returned to the community for further study. Some participants went so far as to run visualization tools on the RSS feed of all the entries as they were being posted, thus returning a feed of the posts live to participants, which allowed study to happen as the day proceeded.

One of the problems we encountered was cleaning up the data after the day. The cleaning up of the data involved four broad steps:

1. To comply with ethics, we had to go through and edit (with the participants) the images posted to make sure the images conformed to the ethics regimen we agreed to.
2. We read and indexed the posts with a uniform set of terms, helping draw out semantic relevance in the data.⁴
3. We converted the XML output from the native WordPress format to a more tractable form. Irrelevant fields were removed and content was unescaped, requiring additional editing toward well-formedness. The final cleaned dataset is being reviewed by project participants with notable experience with markup.
4. Finally, we proofed the entire dataset also deleted empty comments. However, in order to preserve the authenticity of the posts, we did not change the prose of the participants.

3. Crowdsourcing in the Digital Humanities

The Day in the Life of Digital Humanities is a modest example of a collaborative "crowdsourcing" project. It is not the first such project in the humanities. For instance, Suda On Line is an excellent example of how a "crowd" can participate in a larger project.⁵ Reflecting on the level of participation in the Day of DH, we believe that some of the strategies we adopted to encourage participation were successful:

- A participant's required contribution was limited to only one day of posting. We hypothesize that if small, flexible tasks contribute to broad participation.
- We did not assume people would participate. Instead we invited people personally, creating a personal connection before issuing an open call for participation. We believe that the personal human contact makes a real difference in explaining to people why they would want to participate.
- The project was structured as a collaborative publication so that participants could get credit for their work and use the results in their own experiments. We tried to make the idea simple to grasp, which is why we chose the "Day in the Life of" title. The title gives the prospective participant an idea of the level of participation and the results.
- A steady but light feed of updates was maintained through a discussion list. We sent about an e-mail a week to keep in touch as the day approached.

- Human contact and communication are essential at all levels - participants are, after all, volunteering their effort to make the project work. For that reason we had a number of people assigned to answer different types of questions quickly, and we spent some time developing online materials to help explain the project and connect people.
- The technology used by participants was reasonably familiar and worked.

4. Reflections and Theory

What then have we learned about the digital humanities from the project? To some extent the project speaks for itself. The project doesn't provide a short answer to questions about what we do. Instead it provides a wealth of detail and reflections. Nonetheless we do have some conclusions based on readings of the entries:

- Many who do computing in the humanities feel isolated and welcome venues for participating in larger concerns. This project gave some of those isolated a way to feel part of a peer community and to be heard.
- In Humanities research, there is often an inverse relationship between depth and breadth. At their most qualitative and meticulous, humanists may spend years analyzing a short text. To broaden the corpus often necessitates a colder, more mechanical approach to the data. Though perhaps at the expense of structure, the format of Day of DH has resulted in content that is both deep and broad.
- Community projects don't simply document an existing community - to some extent they create it. This is an age-old pattern where a community, in becoming, presents itself as already mature. One participant told us that they were thinking of running something similar at their university as a community-building exercise. While the data is not necessarily an objective representation of what we typically do (if there is such a thing) it is representative of what some of us think about what we do.
- One aim of the Day was to explore the usefulness of autoethnography as a methodology for studying the digital humanities. Nicholas Holt defines autoethnography as a "writing practice [involving] highly personalized accounts where authors draw on their own experiences to extend understanding of a particular discipline or culture".⁶ This reflexive study of the participant-researcher's own role in a greater culture thus has created a dataset far richer and more complex than would have otherwise been available if digital humanists had been given a set of parameters, such as a questionnaire, in which to define themselves.

- Willard McCarty proposes that we think of our practice as one of modeling where we are both modeling as a process of exploration and creating models that represent interpretative process.⁷ This project can be thought of a collaborative modeling of the field where for one day we used some of our own tools and new methods to think about our research in community.

Further observations we leave for you; after all, the point was to leave the community data stream to think about and with.

5. The Second Day of DH

On March 18th, 2010 we plan to run the Second Day of Digital Humanities. This second project will try to address some of the limitations of the first:

- We hope to invite more graduate students to participate. Students appeared resistant to the idea that they had any meaningful contribution to make. One participant, Domenico Fiornante, engaged his students by having them comment on his posts, an approach we will encourage others to do. Another alternative is to encourage students to share a single blog so they don't feel they have to write more than one post.
- We hope to involve more international participants outside Anglophone regions. In particular we hope to involve more Francophone participants in Quebec, but we also plan to invite participants from a broader range of regions and provide them with support early so they feel comfortable posting.
- We hope to find a technology for posting that outputs clean XML without forcing participants to learn markup. The technology will be chosen in conjunction with participants and may be hosted by a participating centre.
- We hope to encourage use of a common set of categories built on those we used for the post-day tagging.
- We plan to better incorporate micro-blogging (Twitter) so that participants could use that technology as an alternative.

6. Conclusion

There are a couple of different lenses that might be appropriate to the discussion of the Day of DH. First, it can be seen as an exercise by the participants and the larger community in building social capital. Bourdieu's work on social capital emphasizes both the actual and potential resources available to the individual through participation in a network.⁸ Coleman focuses on the potential benefits to the individual.⁹ Putnam highlights the value of

social capital to the community, equating community participation with civic virtue.¹⁰ Individuals involved in the DDH have had an opportunity to increase, extend, or consolidate existing social capital through self-revelation within the framework of the day. The DH community in the larger sense has had a moment of opportunity for critical self-reflection.

The second possible lens deals primarily with that possibility for self-reflection. Much as every design can be read as a comment on the act of designing and the discipline of design, or every building as a contribution to the ongoing discussion of architecture, so DDH provides a moment of self-directed reflection on what it means to be a digital humanist in a world where other digital humanists are also active.

Notes

1. Council of Canadian Academies, "The State of Science & Technology in Canada (Summary and Main Findings)", 2006. <http://www.scienceadvice.ca/documents/Summary%20and%20Main%20Findings.pdf>.
2. Ayers, Edward J. "What Does a Professor Do All Day, Anyway?" Lecture given in 1993 at the University of Virginia on receiving the "Teacher of the Year" award at the Fall Convocation and Parents' Weekend. <http://www.virginia.edu/insideuva/textonlyarchive/93-12-10/7.txt>
3. Day in the Life of Digital Humanities wiki. http://tapor.ualberta.ca/taporwiki/index.php/Day_in_the_Life_of_the_Digital_Humanities
4. See http://tapor.ualberta.ca/taporwiki/index.php/Category_Tags for the category tags we used. These were developed iteratively going through the data.
5. Mahoney, Anne. "Tachypaedia Byzantina: The Suda On Line as Collaborative Encyclopedia." *Digital Humanities Quarterly*, V. 3, N. 1. Winter 2009. <http://digitalhumanities.org/dhq/vol/3/1/000025.html>
6. Holt, Nicholas. "Representation, Legitimation, and Autoethnography: An Autoethnographic." *International Journal of Qualitative Methods* 2 (2003): 1-22. Page 1. http://www.ualberta.ca/~iigm/backissues/2_1/pdf/holt.pdf
7. McCarty, Willard. 2005. *Humanities Computing*. Palgrave MacMillan: Basingstoke.
8. Bourdieu Pierre. 1985. «The forms of capital.» In *Handbook of Theory and Research for the Sociology of Education*. Ed. J. G. Richardson. New York: Greenwood, pp. 241-58.
9. Coleman, James S. "Social Capital in the Creation of Human Capital." *American Journal of Sociology*. Volume 94, Number S1: S95. January 1988. Supplement. DOI: 10.1086/228943.
10. Putnam, Robert D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon and Schuster.

Letters, Ideas and Information Technology: Using digital corpora of letters to disclose the circulation of knowledge in the 17th century

Roorda, Dirk

dirk.roorda@dans.knaw.nl

Data Archiving and Networked Services,
Netherlands

Bos, Erik-Jan

erik-jan.bos@phil.uu.nl

Department of Philosophy, Utrecht University,
Netherlands

van den Heuvel, Charles

charles.vandenheuvel@vks.knaw.nl

Virtual Knowledge Studio, Netherlands

1. Circulation of Knowledge and Letters

The scientific revolution of the 17th century was driven by countless discoveries in Europe and overseas in the observatory, in the library, in the workshop and in society at large. There was a dramatic increase in the amount of information, giving rise to new knowledge, theories and world images. But how were new elements of knowledge picked up, processed, disseminated and – ultimately – accepted in broad circles of the educated community? A consortium of universities, research institutes and cultural heritage institutions has started a project called CKCC¹ to meet this research question, building a multidisciplinary collaboratory to analyze a machine-readable and growing corpus of letters of scholars who lived in the 17th-century Dutch Republic. Until the publication of the first scientific journals in the 1660s, letters were by far the most direct and important means of communication between intellectuals. Therefore the 17th-century Republic of Letters offers an ideal case for exploring the answers to this question.

Researchers want to uncover patterns in letters that are indicative for the circulation of knowledge, patterns that reveal the emergence of complex, collective phenomena in modern science. However, they face some fundamental problems with finding such patterns in letters. One cannot know in advance the nature of these patterns, and only few categorical

hypotheses can be tested by simply data mining the letters. Purported patterns cannot be tested against the letters, because the heterogeneous information on which these patterns are based cannot be gleaned from the texts, but need considerable interpretation and contextualization.

Here is a short list of the problems: (i) the letters are not uniformly available; (ii) the 17th century language varieties are not standardised and pose a challenge for language technology; (iii) much interpretation is needed to resolve references to people, places, dates, ideas and instruments; interpretations are complicated by the heterogeneity of annotations; (iv) it is not clear how to set up visualisations of patterns that are really informative to the historian of science. These four types of problems will be used to report on the methodology of the project and on its results so far.

2. Information technology as a humanities' observatory

2.1. Availability of the Letters

CKCC limits itself to the ca. 20,000 letters written by scholars that were active in the Netherlands: René Descartes, Hugo Grotius, Constantijn Huygens, Christiaan Huygens, Caspar Barlaeus, Jan Swammerdam and Anthony van Leeuwenhoek. Modern editions of these correspondences—already published or in an advanced state of production by members of CKCC—form the basis of the digitised texts. The letters, once converted to a minimal TEI format, will then be made available through e-Laborate,² a web-based philological annotation tool that will be transformed into a collaboratory for the history of science and the humanities in general. It serves three purposes: (a) providing scholarly access to the letters; (b) allowing researchers to enrich existing datasets and annotate the letters; (c) using the letters and the input of researchers to visualise patterns meaningful for the circulation of knowledge.

2.2. Use of other datasets

We will incorporate a particular database of (meta)data, the *Catalogus Epistularum Neerlandaricum* (CEN), or the Catalogue of letters in Dutch repositories. It is a relatively old database, already available via Telnet in the early 1990s, before the world wide web came into being.

CEN is an exhaustive database of letters in the collections of five Dutch university libraries, the Royal Library, and four other important libraries. It contains more than 265,000 descriptions of approximately 1,000,000 letters, dating from 1600

until the present day (of which ca. 100,000 from the 17th century). It supplies the following metadata: sender, recipient, place of sending, year, language, repository and shelf mark.

The format in which this database will be made available to the project is to be negotiated with the owner, OCLC.³

Usage of this database will enable us to make assertions about the fraction of the selected letters with respect to the total body of letters. Moreover, it allows us to increase the density of the networks we are interested in, leading to unprecedented research opportunities.

2.3. Language technology

In order to find meaningful patterns in social networks of scholars and in circulation of knowledge language technology is needed. For this, CKCC is cooperating with CLARIN.⁴ The mission of CLARIN is to make language technology interoperable and to make linguistic resources accessible on a European infrastructure, so that all the arts and humanities can make use of it. The Netherlands pillar of CLARIN, CLARIN-NL,⁵ has already obtained funding for constructing such infrastructure, and has issued a call for proposals for adding existing resources to this infrastructure and writing demonstrator services. Aided by expertise provided by CLARIN members, in particular by the University of Lancaster,⁶ CKCC is developing such a demonstrator. A proposal to this end has been accepted by CLARIN-NL. The demonstrator, comprising the correspondences of Grotius, Const. Huygens and Descartes (ca. 15,000 letters in all), is planned to be completed by October 2010. It will perform a time-sensitive keyword extraction, which can be visualised by means of a dynamic word cloud. As the source languages are 17th century Dutch, French and Latin, one needs at least spelling normalisation and harmonisation of keywords across languages.

2.4. Interpretation and Enrichment

References to people, places and times are often implicit and can only be retrieved by studying contextual material or by using secondary sources. Named Entity Recognisers are helpful, but it is not possible to rely on technology alone. In order to get an accurate picture in sufficient resolution, interplay between manual work and automatic tools is needed. The collaboratory based on e-Laborate gives researchers the opportunity to collect their interpretations of the texts, compare them to others and to annotate them with their insights. Over time, the results of this hand/mind work might be

automatically gathered and incorporated in enriched transcriptions of the texts.

2.5. Visualisation

By offering meaningful visualizations of the data, the CKCC will enable humanities researchers in a wider context to use the tools and the results yielded. Not only the relationships between corresponding authors will be made visible in time and space, but CKCC also aims at visualizing the dynamics of knowledge production by focusing on the emergence of themes in scholarly debates and social networks of 17th century natural philosophers.

The dynamic word clouds based on keyword extraction is just a first step. CKCC will subsequently explore several approaches of gathering and visualising meaningful patterns, which are deliberately different in nature. The first approach (a) is a sophistication of keyword extraction, and the second one (b) is based on associations in text. Both methods can be used to evaluate the results of each other.

(a) Concept analysis. This requires considerable more analysis than keyword extraction. For example: the many surface expressions of a concept must be linked into one entity, preferably part of an ontology or thesaurus. Existing subject indices and reference corpora will be used. Visualising the behaviour of concepts over time will yield a good approximation of knowledge circulation.

(b) Associative neural network technology. This is an application of a recent effort, ANNH,⁷ to apply the idea of neural networks to the humanities. This approach enables the automatic comparison of texts, based on the degree of associative similarity. Concepts or themes occurring in the letters could thus be made visible, either by focussing on single terms (e.g. ‘observations’) or word pairs (e.g. ‘soul’ and ‘matter’). Moreover, it is possible to query for letters associated to a given one, and rank the results by the degree of association. The facility to track the circulation of knowledge will then be within reach.

3. An infrastructure for the digital humanities

Infrastructure is of particular interest in view of current developments in Europe as testified by the ESFRI roadmap.⁸ The roadmap funds the preparation for several research infrastructures in the humanities, among which CLARIN, as demonstrated above, is most relevant for the CKCC project. CKCC will take care that the materials residing in the collaboratory can be exported in such a way that it is available on the CLARIN

infrastructure, thus contributing to the much grander vision to have all scholarly letters of the 17th century uniformly available for research, including the results of related work.⁹

In due course, CKCC will not only contribute to the understanding of the circulation of knowledge in the 17th century, but also generate useful technologies for cross-disciplinary collaborations involving data-sharing and data-enrichment in the Humanities. As such, this web-based humanities collaboratory on correspondences is a valuable prototype for possible future research collaborations focusing on large, heterogeneous datasets in the Humanities.

References

- Holthausen, K and Ziche, P.** (2007). 'Neuronale Netze für die Geisteswissenschaften'. *Akademie Aktuell*. **20**: 32-35. http://www.badw.de/aktuell/akademie_aktuell/2007/heft1/10_Holthausen.pdf.

Notes

1. Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic. A Web-based Humanities Collaboratory on Correspondences. Participants: Descartes Centre for the History and Philosophy of the Sciences and the Humanities, Utrecht University; National Library of the Netherlands; Huygens Institute, http://www.huygensinstituut.knaw.nl/_eng/ (accessed 2010-03-19); DANS; Virtual Knowledge Studio. Start date: November 2008. Duration: 4 years. Budget: 1 M€.
2. e-Laborate, <http://www.e-laborate.nl/en/> (accessed 2010-03-19), Huygens Instituut.
3. Online Computer Library Center, <http://www.oclc.org> (accessed 2010-03-19).
4. Common Language Resources and Technology Infrastructure, <http://www.clarin.eu/> (accessed 2010-03-19). CKCC is on the list of supported projects by which CLARIN is reaching out to the humanities, <http://www.clarin.eu/wp3/wp3/wp3-documents/call-for-full-proposals-for-collaboration-with-humanities-and-social-sciences> (accessed 2010-03-19).
5. CLARIN-NL (Netherlands), <http://www.clarin.nl/node/2> (accessed 2010-03-19).
6. University Centre for Computer Corpus Research on Language, <http://ucrel.lancs.ac.uk/> (accessed 2010-03-19), in particular Paul Rayson, <http://www.comp.lancs.ac.uk/~paul/> (accessed 2010-03-19).
7. Associative Neural Networks for the Humanities. An application developed by the Department of Philosophy (P. Ziche, E.-O. Onnasch, E.-J. Bos), Utrecht University, and Dr. Holthausen GmbH, Bocholt. See (Holthausen et al., 2007).
8. European Roadmap for Research Infrastructures, <http://cordis.europa.eu/esfri/roadmap.htm> (accessed 2010-03-19).

9. Mapping the Republic of letters, <http://shc.stanford.edu/collaborations/supported-projects/mapping-republic-letters> (accessed 2010-03-19). Cultures of Knowledge, <http://www.history.ox.ac.uk/cofk/> (accessed 2010-03-19).

Pointless Babble or Enabled Backchannel: Conference Use of Twitter by Digital Humanists

Ross, Claire

claire.ross@ucl.ac.uk

Department of Information Studies, University College London

Terras, Melissa

m.terras@ucl.ac.uk

Department of Information Studies, University College London

Warwick, Claire

c.warwick@ucl.ac.uk

Department of Information Studies, University College London

Welsh, Anne

a.welsh@ucl.ac.uk

Department of Information Studies, University College London

Microblogging, a variant of a blogging which allows users to quickly post short updates to websites such as twitter.com, has recently emerged as a dominant form of information interchange and interaction for academic communities. To date, few studies have been undertaken to make explicit how such technologies are used by and can benefit scholars. This paper considers the use of Twitter as a digital backchannel by the Digital Humanities community, taking as its focus postings to Twitter during three different international 2009 conferences. This paper poses the following question: does the use of a Twitter enabled backchannel enhance the conference experience, collaboration and the co-construction of knowledge, or is it a disruptive, disparaging and a inconsequential tool full of ‘pointless babble’?

Microblogging, with special emphasis on Twitter.com,¹ the most well known microblogging service, is increasingly used as a means of extending commentary and discussion during academic conferences. This digital “backchannel” communication (non-verbal, real-time, communication which does not interrupt a presenter or event, Ynge 1970) is becoming more prevalent at academic conferences, in educational use, and in organizational settings. Frameworks for understanding the role and use of digital backchannel communication, such as that provided by Twitter,

in enabling a participatory conference culture are therefore required.

Formal conference presentations still mainly occur in a traditional setting; a divided space with a ‘front’ area for the speaker and a larger ‘back’ area for the audience. Implying a single focus of attention. There is a growing body of literature describing problems with a traditional conference setting; lack of feedback, nervousness about asking questions and a single speaker paradigm (Anderson et al 2003, Reinhardt et al 2009). The use of a digital backchannel such as Twitter, positioned in contrast with the formal or official conference programme, can address this, providing an irregular, or unofficial means of communication (McCarthy & Boyd, 2005). Backchannel benefits include being able to ask questions, or provide resources and references, changing the dynamics of the room from a one to many transmission to a many to many interaction, without disrupting the main channel communication. Negatives include a cause of distraction, disrespectful content and creating cliques amongst participants (Jacobs & Mcfarlane 2005, McCarthy and Boyd 2005). Nevertheless research consistently shows the digital backchannel as a valuable way for active participation (Kelly 2009) and that it is highly appropriate for use in learning based environments (Reinhardt et al. 2009). Recently microblogging has been adopted by conferences such as DH2009 to act as a backchannel as it allows for the ‘spontaneous co-construction of digital artefacts’ (Costa et al 2008). Such communication usually involves note taking, sharing resources and individuals real time reactions to events.

This paper presents a study that analyses the use of Twitter as a backchannel for academic conferences, focusing on the Digital Humanities community in three different physical conference settings held from June to September 2009. During three key conferences in the academic field (Digital Humanities 2009, That Camp 2009 and Digital Resources in the Arts and Humanities 2009), unofficial Twitter backchannels were established using conference specific hashtags (#dh09, #thatcamp and #drha09, #drha2009¹²) to enable visible commentary and discussion. The resulting corpus of individual “tweets” provides a rich dataset, allowing analysis of the use of Twitter in an academic setting, and specifically presenting how the Digital Humanities community has embraced this microblogging tool.

1. Method

Data from the three conferences was collected by archiving tweets which used the four distinct

conference hashtags. (These hashtags were used prior to and after the conferences, and have been reused by other conferences, therefore the corpus was limited to tweets posted during the span of each conference). This provided a data set of 4574 tweets from 326 distinct Twitter users, resulting in a corpus of 77308 tokens, which were analysed using various quantitative and qualitative methods which allowed us to understand and categorize the resulting corpus effectively.

Quantitative measures were used such as identifying prominent tweeters, analysing the frequency of conversations between users and the frequency of reposting messages (“retweeting”), and the differing use of Twitter at the three separate events. Text analysis tools were also used to interrogate the corpus.

Tweets were then categorized qualitatively using open coded content analysis where each post was read and categorized, determining the apparent intention of each twitter post. It was necessary to develop our own categories: although Java et al (2007) present a brief taxonomy of Twitter user intentions (daily chatter, conversations, sharing information and reporting news) they are based on general Twitter use and were too imprecise for our needs. Ebner (2008) discovered four major categories in his study of the use of Twitter during the keynote presentation at the Ed-Media 2008 conference, but this is a small study limited to 54 posts made by ten distinct users, whereas the DH conferences involved a much larger user population. Through our analysis, Tweets were divided into seven categories: comments on presentations; sharing resources; discussions and conversations; jotting down notes; establishing an online presence; and asking organizational questions. Tweets which were highly ambiguous were placed in an Unknown category.

2. Findings

Conference hashtagged Twitter activity does not constitute a single distributed conversation but, rather multiple monologues and a few intermittent, loosely joined dialogues between users. The majority of the activity was original tweeting (93%): only 6.7% were re-tweets (RT) of others’ ideas or comments. The real time interchange and speed of review of shared ideas appears to create a context of users offering ideas and summaries and not spreading the ideas of others verbatim. 45% of the tweets during the conference proceedings included direct references to others’ Twitter IDs, using the ‘@’ sign, as the source of a quote, object of a reply or debate. This implies a form of collaborative writing activity, driving a conference community of practice (Wenger 1998)

who are involved in shared meaning making and the co-construction of knowledge (McNely 2009). However, the content of the tweets indicate that the discussion was between a few Twitter users rather than mass collaboration and was not necessarily focused on conference content.

Jacob and Mcfarlane (2005) discuss polarization in digital backchannels, highlighting a conflict between an inclusive and participatory conference culture and a fragmentation of conference participants into cliques only intermittently engaged with the main presentations. This, in some instances seems to be the case during the Digital Humanities conferences, suggesting that newer members of the discipline or newer uses to Twitter may be excluded from the discussion. This also raises the question about official and unofficial backchannels. When communication is digitally mediated, backchannels may not be obvious. That is, even if participants know who else is participating in an interaction, this doesn't guarantee (as it does in the front channel) that it is an accessible backchannel. Therefore by its nature an unofficial backchannel does not enable active participation. Further research is currently being undertaken on this corpus, which will be presented fully in the paper.

Most tweets in the corpus fell into the category of jotting down notes, triggered predominately by the front channel presentation, suggesting that participants are sharing experiences and to a degree co-constructing knowledge. What is surprising is the lack of direct commentary on presentations. Although Reinhardt et al (2009) argue that Twitter enables thematic debates and offers a digital backchannel for further discussion and commentary, the tweet data suggests that this does not appear to have happened to a significant extent at the digital humanities conferences. This raises the question of whether a Twitter enabled backchannel promotes more of an opportunity for users to establish an online presence and enhance their digital identity rather than encouraging a participatory conference culture.

3. Conclusion

This study of digital humanities conference tweets provides an insight into the Digital Humanities community of practice. The Twitter enabled backchannel constitutes a complex multidirectional discursive space in which the conference participants make notes, share resources, hold discussions and ask questions as well as establishing a clear individual online presence. While determining individual user intentions in Twitter in a conference setting is challenging, it is possible to describe broad behavioral trends of tweeting during Digital Humanities conferences. The predominance of note

taking suggests that the DH community could be classed as social reporters, commenting on the conference presentations for outsiders, rather than collaborating during the conference. There was also a tendency for a small group of users to produce the majority of tweets, interacting with each other about other matters. This suggests the small friendly nature of the DH researcher community, but may also be somewhat intimidating for those new to the field or conference. The Twitter enabled backchannel thus raises some interesting questions about the nature of conference participation and whether or not it is helped or hindered by a digital backchannel. Rather than pointless babble, the twitter record produced at each conference provides important evidence regarding how Digital Humanities, as a community of practice, functions and interacts.

References

- Anderson, R.J., Anderson, R., Vandergrift, T., Wolfman, S., Yasuhara, K.** (2003). 'Promoting Interaction in Large Classes with Computer Mediated Feedback'. *Designing for Change in Networked Learning Environments. Proceedings of CSCC*. Bergen, pp. 119-123.
- Barnes, S.J., Böhringer, M.** (2009). 'Continuance Usage Intention in Microblogging services: The Case of Twitter'. *Proceedings of the 17th European Conference on Information Systems (ECIS)*. <http://www.ecis2009.it/papers/ecis2009-0164.pdf>.
- Costa, C., Beham, G., Reinhardt, W., Sillaots, M.** (2008). 'Microblogging In Technology Enhanced Learning: A Use-Case Inspection of PPE Summer School 2008'. *Microblogging in technology enhanced learning conferences: A use case Inspection Workshop at the European Conference on Technology Enhanced Learning*. Maastricht, the Netherlands, 2008.
- Ebner, M.** (2009). 'Introducing Live Microblogging: How Single Presentations Can Be Enhanced by the Mass'. *Journal of Research in Innovative Teaching. Volume 2, Issue 1*: 91-100.
- Ebner, M., Schiefner, M.** (2008). 'Microblogging - more than fun?'. *Proceedings of IADIS Mobile Learning Conference*. Algarve, Portugal, 2008, pp. 155-159.
- Jacobs, N., Mcfarlane, A.** (2005). 'Conferences as Learning Communities: Some early lessons in using 'back-channel' Technologies at an Academic Conference - Distributed Intelligence or Divided Attention'. *Journal of Computer Assisted Learning. 21 (5)*: 317-329.

Java, A., Song, X., Finin, T., Tseng, B. (2007). 'Why we twitter: understanding microblogging usage and communities'. *International Conference on Knowledge Discovery and Data Mining Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. New York, 2007.

Karger, D. R., Quan, D (2005). 'What would it mean to blog on the semantic Web?'. *Web Semantics: Sciences, Services and Agents on the World Wide Web*. 3: 147-157.

Kelly, B. (2009). (*TwitterFall*) *Your're My Wonder Wall*. *UK Web Focus: Reflections on the Web and Web 2.0*. <http://ukwebfocus.wordpress.com/2009/04/20/twitterfall-youre-my-wonder-wall/>.

Krishnamurthy, B., Gill, P., Arlitt, M. (2008). 'A few Chirps about Twitter'. *WOSP 08: Proceedings of the first workshop on online social networks*.

McCarthy, J.F., Boyd, D. (2005). 'Digital backchannels in shared physical spaces: Experiences at an academic Conference'. *Proceedings of the Conference on Human Factors in Computing Systems*.

McFedries, P. (2007). 'All a-twitter'. *IEE Spectrum*. 84.

McNely, B.J. (2009). 'Backchannel Persistence and Collaborative Meaning Making'. *Proceedings of the 27th ACM international conference on Design of communication*. Bloomington, Indiana.

Reinhardt, W., Ebner M., Beham G., Costa, C. (2009). 'How People are using Twitter during Conferences'. *Creativity and Innovation Competencies on the Web. Proceeding of 5. EduMedia conference*.

Wenger, E. (1998). *Communities of practice: learning, meaning, and identity*. Cambridge University Press .

Ynge, V. (1970). 'On Getting a Word in Edgewise'. *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*. Pp. 567-577.

Notes

1. Twitter was created by a San Francisco based privately funded startup and launched publicly in August 2006. <http://twitter.com/about>
2. The community aspect of Twitter means that participants self organize, instigating tags themselves, hence the participants of Digital Resources in the Arts and Humanities used two different hashtags to discuss the conference depending on the twitter user.

The State of Non-Traditional Authorship Attribution Studies – 2010: Some Problems and Solutions

Rudman, Joseph
jr2o@heps.phys.cmu.edu
Carnegie Mellon University, USA

In 1997, at the ACH-ALLC'97 conference at Queen's University, there was a session presented by R. Harald Baayen, David I. Holmes, Joe Rudman, and Fiona J. Tweedie, "The State of Authorship Attribution Studies: (1) The History and the Scope; (2) The Problems – Towards Credibility and Validity." Thirteen years have passed and well over 600 studies and papers dealing with non-traditional authorship attribution have been promulgated since that session.

This paper looks back at that session, a subsequent article published by Rudman in *Computers and the Humanities*, "The State of Authorship Attribution Studies: Some Problems and Solutions," and the more than 600 new publications. There are still major problems in the "science" on non-traditional authorship attribution. This paper goes on to assess the present state of the field – its successes, failures, and prospects.

1. Successes

It has been an exciting thirteen years with many advances. Each of the following (not a complete list) will be discussed:

1. Arguably, the most significant development in the field is the large contingent of computer scientists that have brought their perspectives to the table – led by Shlomo Argamon, Moshe Kopple, and a host of others.
2. The Dimacs Working Group on Developing Community.
3. Sir Brian Vickers' London Authorship Forum.
4. John Burrows' Busa Award.
5. Forensic Linguistics.
6. Successful studies such as Foster's *Primary Colors* work.

7. The continuing advances of practitioners such as John Burrows, David Hoover, Matthew Jockers, David Holmes, and others.
8. John Nerbonne's reissue of Mosteller and Wallace's *Applied Bayesian and Classical Inference: The Case of "The Federalist Papers."*
9. Patrick Juola's "Ad Hoc Authorship Attribution Competition." and his NSF funded JGAAP project.
10. The PAN Workshops. Uncovering Plagiarism, Authorship, and Social Software Misuse.

2. Acceptance

Contrary to what many practitioners of the non-traditional proclaim, there is not wide-spread acceptance of the field.

There have been many high profile problems with the concomitant negative publicity, e.g.:

1. Foster's misattribution of *A Funeral Elegie*
2. Foster's misattribution of the Jon Benét ransom note
3. Burrows' attribution then de-attribution of "A Vision"
4. The continuing bashing of Morton's CUSUM

Burrows' shift is something that every good scientist should do – search for errors or improvements in their experimental methodology and self correct.

3. Failures and Shortcomings

After thirteen years of increasing activity, there is still no consensus as to correct methodology or technique. Most authorship studies are still governed by expediency, e.g.:

- The texts are not the correct ones but they were available
- The controls are not complete but it would have taken too long to obtain the correct ones

The "umbrella" problem remains – most non-traditional authorship practitioners do not understand what constitutes a valid study.

Problems in the following areas will be explicated and solutions proposed:

- Knowledge of the Field (i.e. the Bibliography) – The fact that there have been so many authorship studies is good -- the fact that they have been published in over 90 different journals makes a complete literature search time consuming and difficult which is not good. To make things even more difficult, add to this the more than 14 books, 22 chapters in books, the 80 conference papers,

the 10 reports, 22 dissertations, 9 newspaper articles, the 10 on-line self published papers, 4 encyclopedia entries.

- Reproducibility – verification
- The Experimental Plan
- The Primary Data – This is a major problem that is almost universally side-stepped.
- Style markers – Function words, n-grams, etc.
- Cross Validation – necessary but not sufficient
- The Control Groups – Genre, gender, time frame, etc.
- The Statistics – A range of techniques will be discussed – e.g. Neural Nets, SVM's, Sequence Kernels, Nave Bayes
- The Presentation – visualization

4. Conclusion

In conclusion, there is a discussion of our role as gatekeepers:

- Rudman's caution that attribution studies on the *Historia Augusta* are an exercise in futility.
- Hoover and Argamon's modification and clarification of Burrows' Delta.
- Rudman's "Ripost" of Burrows' "History of Ophelia."
- Should we oppose patents such as Chaski's?
- The Daubert triangle.

References

Argamon, Shlomo, et al. (2003). 'Gender, Genre, and Writing Style in Formal Written Texts'. *Text*. **23.3**: 321-346.

Baayen, Harald, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. (2002). 'An Experiment in Authorship Attribution'. *JADT 2002:6es Journées Internationales d'Analyse Statistique des Données Textuelles*.

Brennan, Michael, and Rachel Greenstadt. *Practical Attacks Against Authorship Attribution Techniques*. <http://www.cs.drexel.edu/greenie/brennan-paper.pdf> (accessed July 14, 2009).

Burrows, John (2007). 'Sarah and Henry Fielding and the Authorship of The History of Ophelia: A Computational Analysis'. *Script & Print*. **30.2**: 69-92.

Chung, Cindy, and James PenneBaker (2007). 'The Psychological Functions of Function Words'.

Social Communication. K. Fiedler (ed.). New York: Psychology Press, pp. 343-359.

Feiguina, Ol'ga, and Graeme Hirst (2007). 'Authorship Attribution for Small Texts: Literary and Forensic Experiments'. *Proceedings of SIGIR '07 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*. Amsterdam.

Foster, Donald W. (February 26, 1996). *Primary Culprit: An Analysis of a Novel of Politics*. New York, pp. 50-57.

Khosmood, Foaad, and Robert Levinson (2006). 'Toward Unification of Source Attribution Processes and Techniques'. *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*. Dalian, pp. 4551-4556.

Love, Harold (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

Niederkorn, William S. (20 June 2002). *The New York Times*, B1, B5.

Ramyaa, Congzhou, and Khaled Rasheed (2004). 'Using Machine Learning Techniques for Stylometry'. *International Conference on Machine Learning (MLMTA'2004)*. Las Vegas.

Rudman, Joseph (2007). 'Sarah and Henry Fielding and the Authorship of "The History of Ophelia": A Ripost'. *Script & Print*. **31.3**: 147-163.

Solon, Lawrence M., and Peter M. Tiersma (2005). *Speaking of Crime*. Chicago: The University of Chicago Press.

Stamatatos, Efstatios, Nikos Fakotakis, and George Kokkinakis (2001). 'Automatic Text Categorization in Terms of Genre and Author'. *Computational Linguistics*. **26.4**: 471-495.

Stein, Benno, et al. (eds.) (2009). *PAN'09*.

Tambouratzis, George (2001). 'Assessing the Effectiveness of Feature Groups in Author Recognition Tasks with the SOM Model'. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*. **36.2**: 249-259.

Van Halteren, Hans (2004). 'Linguistic Profiling for Authorship Recognition and Verification'. *42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona.

Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?

Rybicki, Jan

jkrybicki@gmail.com

Pedagogical University, Krakow, Poland

Eder, Maciej

maciej_eder@poczta.onet.pl

Pedagogical University, Krakow, Poland

In 2007, John Burrows identified three regions in word frequency lists of corpora in authorship attribution and stylometry. The first of these regions consists of the most frequent words, for which his Delta has become the best-known method of study. This is evidenced by a varied body of research with interesting modifications of the method (e.g. Argamon 2008; Hoover 2004, 2004a). At the other end of the frequency list, Iota deals with the lowest-frequency words, while "the large area between the extremes of ubiquity and rarity" (Burrows, 2007) is now the target of many studies employing Zeta (e.g. Craig, Kinney, 2009; Hoover, 2007).

Due to the popularity of the three methods it was only a matter of time before Delta (and, to a lesser extent, Zeta and Iota) were applied to texts in languages other than Modern English: Middle Dutch (Dalen-Oskam, Zundert, 2007), Old English (García, Martín 2007) and Polish (Eder, Rybicki 2009). Delta has also been used in translation-oriented papers, including Burrows's own work on Juvenal (Burrows, 2002) and Rybicki's attempts at translator attribution (2009).

It has been generally - and mainly empirically - assumed that the use of methods relying on the most frequent words in a corpus should work just as well in other languages as it did in English; this question was approached in any detail only very recently (Juola, 2009). We could not fail to observe that its success rates in Polish, although still high, fell somewhat short of its guessing rate in English (Rybicki 2009a). Also, the already-quoted study by Rybicki (2009) seemed to suggest that, in a corpus of translated literary texts, Delta was much better at recognising the author of the original than the translator. This justified a more in-depth look at the workings of Burrows's method both in its "original" English and in a variety of other languages.

1. Methods

In this study, a single major modification has been applied to the usual Delta process. Each analysis was made for the top 50-5000 most frequent words in the corpus - but then the 50 most frequent words would be omitted and the next 50-5000 words taken for analysis; then the first 100 most frequent words would be omitted, and so on. This was done with a single R script written by Eder; the script produced word frequency tables, calculated Delta and produced "heatmap" graphs of Delta's success rate for each of the frequency list intervals, showing the best combinations of initial word position in wordlist and size of window, including variations of pronoun deletion and culling parameters. Thus, in the resulting heatmap graphs, the horizontal axis presents the size of each wordlist used for one set of Delta calculations; the vertical axis shows how many of the most frequent words were omitted. Each of the runs of the script produced an average of ca. 3000 Delta iterations.

2. Material

The project included the following corpora (used separately); each contained a similar number of texts to be attributed.

Code	Language	Texts	Attribution
E1	English	65 novels from Swift to Conrad	Author
E2	English	32 epic poems from Milton to Tennyson	Author
E3	English	35 translations of Sienkiewicz's novels	Translator
P1	Polish	69 19 th - and early 20 th -century novels from Kraszewski to Øeromski	Author
P2	Polish	95 translations of 19 th - and 20 th -century novels from Balzac to Eco	Author
P3	Polish	95 translations of 19 th - and 20 th -century novels from Balzac to Eco	Translator
F1	French	71 19 th - and 20 th -century novels from Voltaire to Gide	Author
L1	Latin	94 prose texts from Cicero to Gellius	Author
L2	Latin	28 hexameter poems from Lucretius to Jacopo Sannazaro	Author
G1	German	66 literary texts from Goethe to Thomas Mann	Author
H1	Hungarian	64 novels from Kemény to Bródy	Author
I1	Italian	77 novels from Manzoni to D'Annunzio	Author

3. Results

The English novel corpus (E1, Fig. 1) was the one with the best attributions for all available sample sizes starting at the top of the reference corpus word frequency list; it was equally easy to attribute even if the first 2000 most frequent words were omitted in the analysis - or even the first 3000 for longer samples. The English epic poems (E2, Fig. 2) had their area of best attributive success removed away from the top of the word frequency list, into the 1000th-2000th most-frequent-word region. Some successful attributions could also be made with a variety of wordlists around the 2000 mark, starting at the 1st most frequent word.

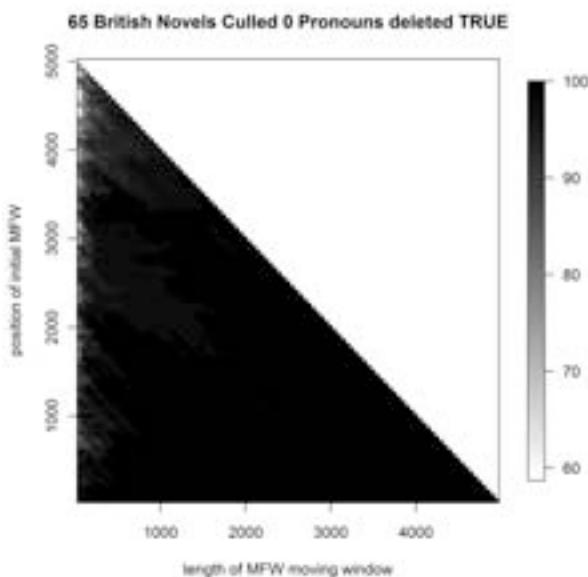


Figure 1. Heatmap of 65 English novels (percentage of correct attributions). Colour coding is from low (white) to high (black)

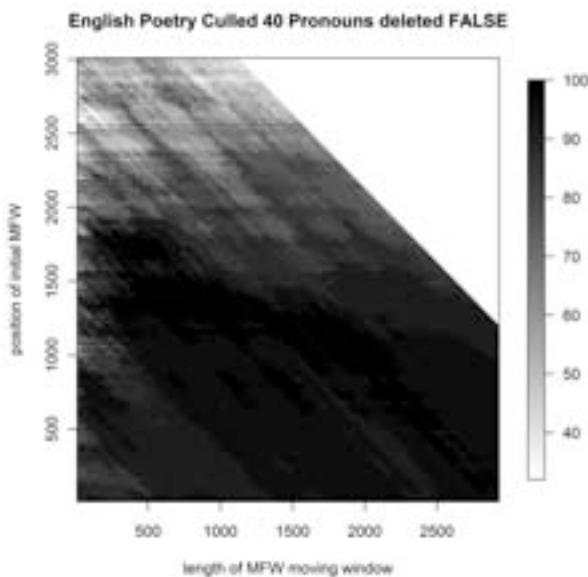


Figure 2. Heatmap of 32 English epic poems

The final "specialist" corpus in the English section of the project - 32 works by Polish novelist Henryk Sienkiewicz, translated into English by a number of translators (Fig. 3) - showed Delta's expected problems in translator attribution; however, for a variety of culling/pronoun deletion parameters, a small yet fairly consistent hotspot would appear for small samples if the first 2000-3000 words were deleted from the frequency wordlist. The first Polish corpus, that of 69 19th- and early 20th-century classic Polish novels (P1, Fig. 4), showed marked improvement in Delta success rate when the wordlist taken for attribution started at some 450 words down the frequency list; the most successful sample sizes were relatively small: no more than 1200 words long.

When the corpus of Polish translations was studied for original authorship (P2, Fig. 5), the results were quite accurate for many sample sizes up to 1800 from the very top of the frequency list. Delta was equally successful for samples of up to 1400 words beyond the 800th most-frequent-word mark. The same corpus yielded lower attribution success when studied for translator recognition (P3, Fig. 6). In fact, it resembled somewhat the graph for Polish classics: a small range of passable attributions, usually for samples below 1000, and usually better when starting a hundred or so words down the frequency rank list.

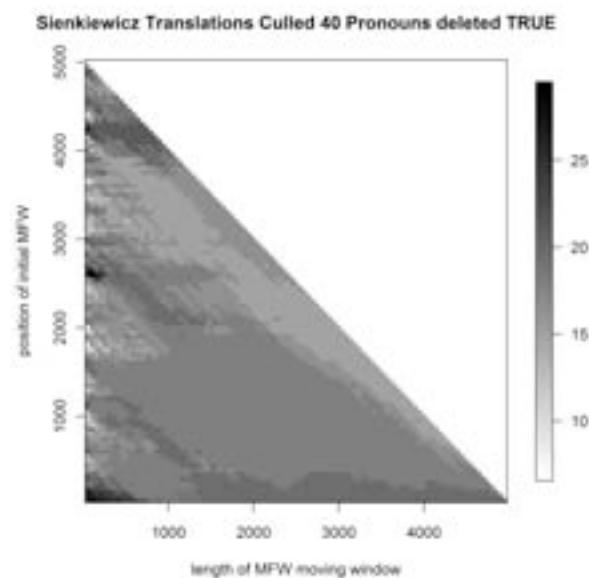


Figure 3. Heatmap of 35 English translations of Sienkiewicz's works

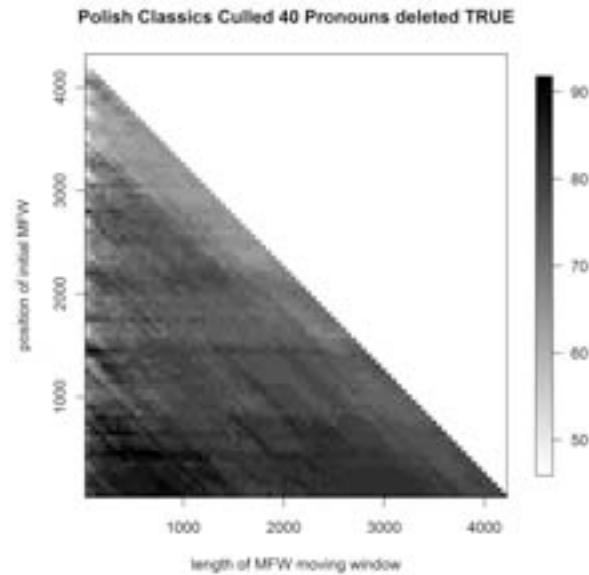


Figure 4. Heatmap of 69 Polish novel classics

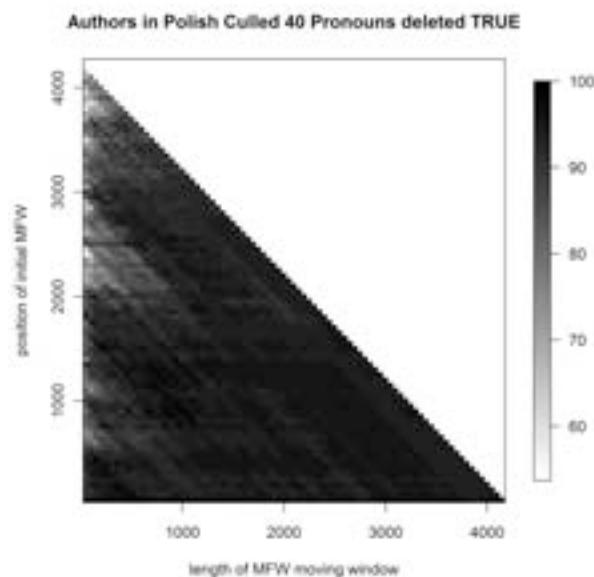


Figure 5. Heatmap of 95 Polish translations from Balzac to Eco (authorship attribution)

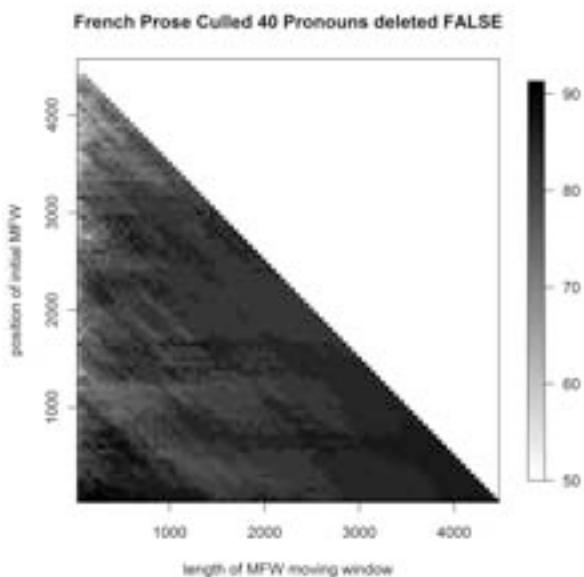


Figure 7. Heatmap of 71 French novels

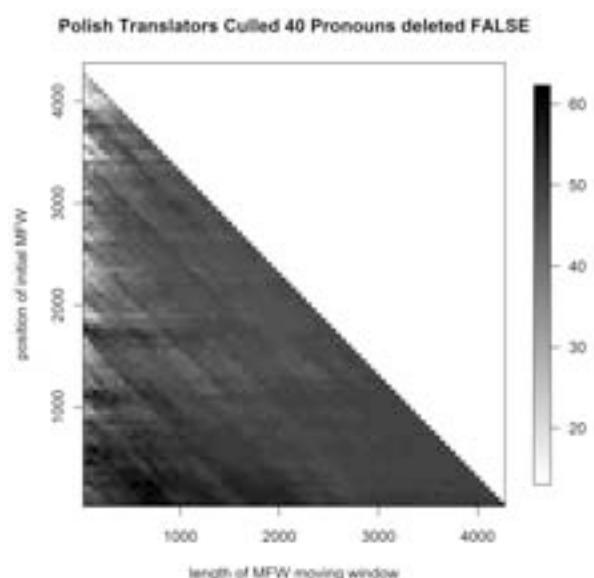


Figure 6. Heatmap of 95 Polish translations from Balzac to Eco (translator attribution)

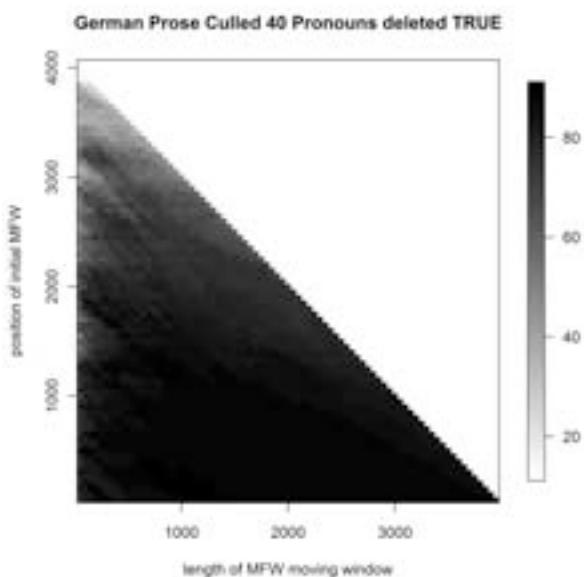


Figure 8. Heatmap of 66 German texts

The French corpus proved almost equally difficult (F1, Fig. 7): Delta was very successful mainly for small-sized samples from the top of the overall frequency wordlist. In contrast, the graph for the German corpus (G1, Fig. 8) presented a success rate akin to that for the English novels, with a consistently high correct attribution in most of the studied spectrum of sample size and for samples beginning anywhere between the 1st and the 1000th word in the corpus frequency list.

Of the two Latin corpora, the prose texts (L1, Fig. 9) could serve as excellent evidence for a minimalist approach in authorship attribution based on most frequent words, as the best (if not perfect) results were obtained by staying close to the axis intersection point: no more than 750 words, taken no further than from the 50th place on the frequency rank list.

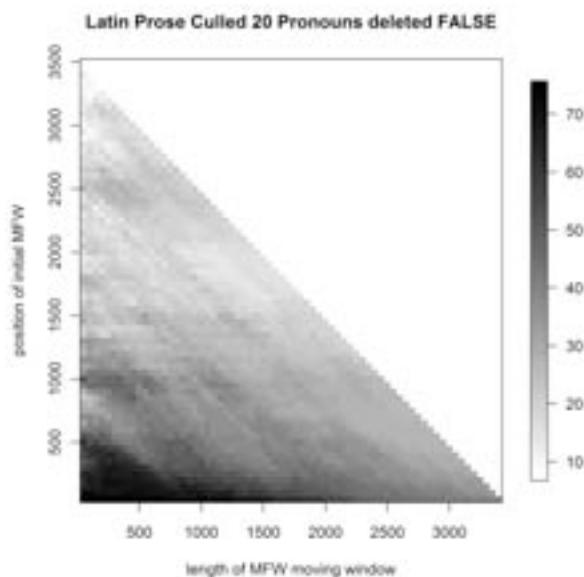


Figure 9. Heatmap of 94 Latin prose texts

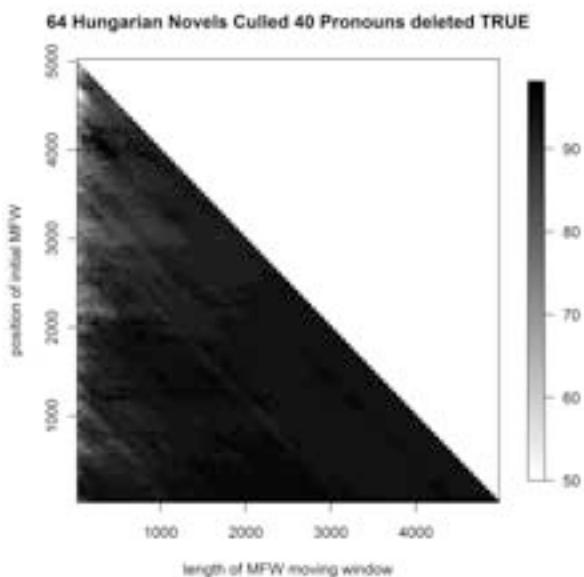


Figure 11. Heatmap of 64 Hungarian novels

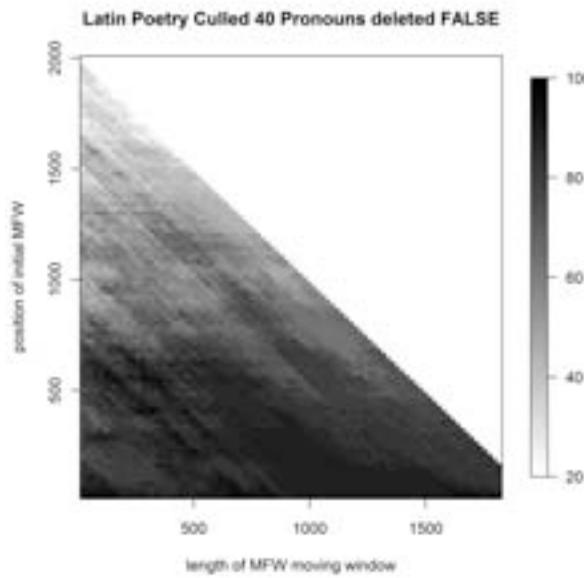


Figure 10. Heatmap of 28 Latin hexameter poems

The other Latin corpus, that of hexameter poetry (L2, Fig. 10), paints a much more heterogeneous picture: Delta was only successful for top words from the frequency list at rare small (150), medium (700) and large (1700) window sizes, and for a few isolated places around the 1000/1000 intersection point in the graph.

The corpus of 19th-century Hungarian novels (H1, Fig. 11) exhibited good success for much of the studied spectrum and an interesting hotspot of short samples at ca. 4000 words from the top of the word frequency list.

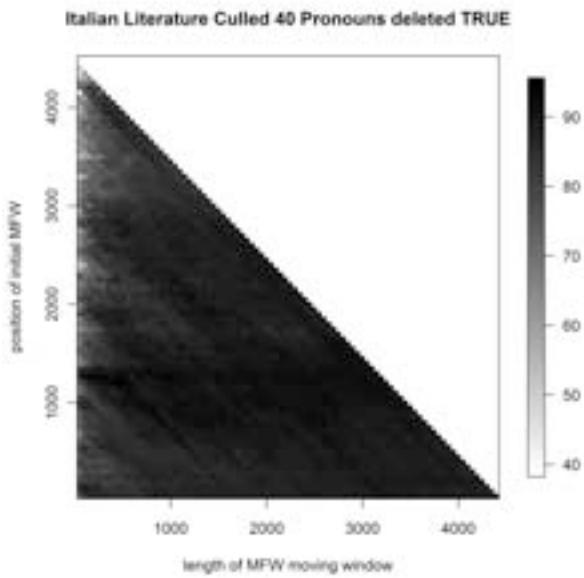


Figure 12. Heatmap of 77 Italian novels

With the Italian novels (I, Fig. 12), Delta was at its best for a broad variety of sample sizes, but only when some 1000 most frequent words were eliminated from the reference corpus.

4. Conclusions

1. Standard Delta (i.e. applied to the most frequent words) provides the best results for authorial attribution in English and German prose.
2. The same procedures still yield acceptable results in other languages and in translator attribution. The success here can be improved by manipulating the number of words taken for analysis and by selecting the reference wordlists at various

- distances from the top of the overall frequency rank list.
3. The differences in attributive success could be partially explained by the differences in the degree of inflection/agglutination of the languages studied, the strongest evidence of this being the relatively highest success rate in English and German.
-

References

- Argamon, S.** (2008). 'Interpreting Burrows's Delta: Geometric and Probabilistic Foundations'. *Literary and Linguistic Computing*. **23(2)**: 131-147.
- Burrows, J.F.** (1987). *Computation into Criticism: a Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.
- Burrows, J.F.** (2007). 'All the Way Through: Testing for Authorship in Different Frequency Strata'. *Literary and Linguistic Computing*. **22(1)**: 27-48.
- Burrows, J.F.** (2002). 'The Englishing of Juvenal: Computational Stylistics and Translated Texts'. *Style*. **36**: 677-99.
- Burrows, J.F.** (2002a). 'Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship'. *Literary and Linguistic Computing*. **17**: 267-287.
- Hoover, D.L.** (2003). 'Frequent Collocations and Authorial Style'. *Literary and Linguistic Computing*. **18**: 261-286.
- Hoover, D.L.** (2004). 'Testing Burrows's Delta'. *Literary and Linguistic Computing*. **19**: 453-475.
- Hoover, D.L.** (2004a). 'Delta Prime?'. *Literary and Linguistic Computing*. **19**: 477-495.
- Hoover, D.L.** (2007). 'Corpus Stylistics, Stylometry, and the Styles of Henry James'. *Style*. **41(2)**: 174-203.
- Rybicki, J.** (2009). 'Translation and Delta Revisited: When We Read Translations, Is It the Author or the Translator that We Really Read?'. *Digital Humanities*. College Park, 2009.
- Rybicki, J.** (2009a). 'Liczenie krasnoludków. Trochę inaczej o polskich przekładach trylogii Tolkiena'. *Po co ludziom krasnoludki?*. Warszawa, 2009.
- Craig, H., Kinney, A.F. (eds.)** (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.
- 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.
- Eder, M., Rybicki, J.** (2009). 'PCA, Delta, JGAAP and Polish Poetry of the 16th and the 17th Centuries: Who Wrote the Dirty Stuff?'. *Digital Humanities*. College Park, 2009.
- Garcia, A.M., Martin, J.C.** (2007). 'Function Words in Authorship Attribution Studies'. *Literary and Linguistic Computing*. **22**: 49-66.
- 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*. **22**: 465-491.
- Mosteller, F., Wallace, D.L.** (1964) (2007). *Inference and Disputed Authorship: The Federalist*. CSLI Publications.
-

Notes

1. Reprinted with a new introduction by John Nerbonne

Reading Darwin Between the Lines: A Computer-Assisted Analysis of the Concept of Evolution in *The Origin of Species*

Sainte-Marie, Maxime B.

msaintemarie@gmail.com

Université du Québec à Montréal, Canada

Meunier, Jean-Guy

meunier.jean-guy@uqam.ca

Université du Québec à Montréal, Canada

Payette, Nicolas

nicolaspayette@gmail.com

Université du Québec à Montréal, Canada

Chartier, Jean-François

chartier.jf@gmail.com

Université du Québec à Montréal, Canada

Whereas Darwin is nowadays considered the founder of the modern theory of evolution, he wasn't the first to use this word in a biological context: indeed, the word "evolution" already had two distinct biological uses at the time the *Origin of Species* was first published (Bowler, 2003; Huxley, 1897): "initially, to refer to the particular embryological theory of preformationism; and later, to characterize the general belief that species have descended from one another over time" (Richards, 1998: 4).

Deriving from the Latin *evolutio*, which refers to the scroll-like act of unfolding or unrolling, the word «evolution» was first used in biology to refer to the development of the embryo, mainly through the formulation, promulgation, and justification of preformationist and epigenetical theories. Embryological evolution would receive its fullest, most modern experimental and theoretical account in the works of Karl Ernst von Baer: characterizing embryological development as a gradual differentiation process leading from homogeneous matter to the production of heterogeneity and complexity of structure, von Baer would usually use the word *Entwickelung* to refer to this dynamic phenomenon, often followed by the Latin *evolutio* in parentheses. The ground-breaking importance of von Baer's work, as well as its diffusion in the scientific community through numerous translations, commentaries, and appropriations, significantly contributed to consecrate the embryological use of the word "evolution".

As for the use of the word evolution to describe specific development, its emergence is closely tied to Lamarckism: even though Lamarck never used the word 'evolution' himself to refer to the transformation of species over time and generations, his commentators, detractors, readers and followers often did however, thus contributing to the semantic alteration of the term. Indeed, "by the 1830s, the word "evolution" had shifted 180 degrees from its original employment and was used to refer indifferently to both embryological and species progression" (Richards, 1992: 15): Étienne Renaud Serres used the expression *théorie des evolutions* in his 1827 article *Théories des formations organiques* to refer both "to the recapitulatory *métamorphoses* of organic parts in the individual and the parallel changes one sees in moving (intellectually) from one family of animals to another and from one class to another" (Richards, 1992: 69); von Baer, in rejecting the possibility of transmutation and the popular idea that embryological development recapitulates the progression of the species, used the word «evolution» to refer to both processes; in England, naturalists such as Charles Lyell, Joseph Henry Green, Robert Grant and Richard Owen also used the word "evolution" to both comment and reject Lamarckism (Bowler, 2003; Richards, 1993).

While this dual usage of the word and its most common synonyms at the time (transformation, development, transmutation...) has been confirmed in the works of the most important biologists and naturalists of the first half of the 19th century, little is known about Darwin's own stance on this matter: did he or not use the word 'evolution' or any other word to refer both to embryological and specific development? This question, however crucial it may appear, proves very difficult to answer: while the *Origin of Species* is generally considered as the birth document of the theory of evolution, studies on and around this book often overlook the fact that the word itself is rarely used by Darwin, the sole and slight exception being the sixth and last edition (1872) of the work.

1 st Edition (1)	<i>evolved</i> : XV (490)
2 nd Edition (1)	<i>evolved</i> : XV (490)
3 rd Edition (1)	<i>evolved</i> : XV (525)
4 th Edition (1)	<i>evolved</i> : XV (577)
5 th Edition (2)	<i>evolved</i> : XV (573), XV (579)
6 th Edition (14)	<i>evolution</i> : (VII:201(2), 202), VIII (215), X (282), XV (424 (3)) <i>evolve</i> : VII (191) <i>evolved</i> : VII (191, 202(2)), XV (425, 429)

Occurrences of "evolution", "evolve", and "evolved" in the *Origin of Species*

This lexical scarcity doesn't necessarily mean however that the concept of evolution isn't present

elsewhere in the text, where the words 'evolution', 'evolved', and 'evolve' don't appear. According to distributional semantics theory, meaning can be more easily stated as a property of word combinations than of words *per se*: in every sentence and paragraph, each word brings its own constraints to the whole, reduces the sets of possible words that could fit with it, therefore increasing the total information conveyed and structuring the semantic dimension of each word thus combined. In short, this theory holds that "similarities and patterning among the co-occurrence likelihoods of various words correlate with similarities and patterning in their types of meaning" (Harris, 1991: 341). In this sense, if concepts are thought of as networks of such meaning-bearing word combinations, then, conceptual structures can determine the semantic dimension of a text without being properly lexicalized; in other words, such considerations, while emphasizing the distinction between the semantic associations of specific concepts and their embodiment in natural language, also seem to imply the possibility of "reading between the lines", that is, of identifying and analyzing concepts on the sole basis of their relations with other words and concepts and independently of any proper designation.

In view of this, the fact that the word "evolution" itself is rarely found in the sixth edition of the *Origin of Species* doesn't necessarily imply that the lexical and inferential network it refers to and that constitutes its conceptual dimension isn't present elsewhere in the text and can't be studied in its stead. In this sense, taking into account word combinations similar to those where the word 'evolution' occurs instead of focusing solely on the latter might be the most reliable way to determine whether or not Darwin's concept of evolution in the *Origin of Species* refers to both embryological and specific development, like most biological theories of the same period. However, dealing with word combinations manually might prove difficult, if not impossible. In light of this, a new computer-assisted conceptual analysis tool has been developed by the LANCI laboratory, one which aims to "read Darwin between the lines", that is to identify where the author "conceptually" refers to evolution, regardless of the presence or the absence of the word itself.

Theoretically speaking, this new approach is based on two fundamental assumptions: 1) The inferential nature and dimension of a concept are linguistically expressed in a differentiated, contextualized and regularized manner; 2) these regularities and patterns can be identified or distinguished using algorithmic, iterative and automatic clustering methods. Concretely, the algorithm aims at "digging deeper into data" by means of an iterative clustering process. Following an initial clustering of the

analyzed corpus (in this case, the 974 paragraphs of the sixth edition of the *Origin of Species*), the iterative concordance clustering process starts by retrieving the most characteristic word of each cluster containing the word(s) to be analyzed, that is, the word that has the highest TF.IDF rating (Term Frequency – Inverted Document Frequency) for each of these clusters. Then, the concordance of each of these characteristic words is extracted from the corpus, and the same process of clustering, cluster selection, TF.IDF rating and ranking, word selection and concordance extraction is performed on each of those new concordances, until no new characteristic word is found or no more clusters containing the word(s) to be analyzed are found.

1. Concordance extraction:	For each cluster containing the word(s) to be analyzed, extract the concordance of the highest-TF.IDF-ranked word.
2. Concordance clustering:	For each previously unselected word, proceed to the clustering of its concordance.
3. Iteration:	Return to step 1, unless 1) no new highest-TF.IDF-ranked word is found, or 2) no clusters containing the word(s) to be analyzed are found.

Iterative Concordance Clustering Algorithm

In order to identify the principal lexical constituents of the concept of evolution and determine whether or not this underlying conceptual structure includes references to both embryological and specific processes, two different extraction procedures were made: the first one only aimed at the word "evolution", while the second one also added "evolve" and "evolved". Figures 1 and 2 show the results of the two analyses.

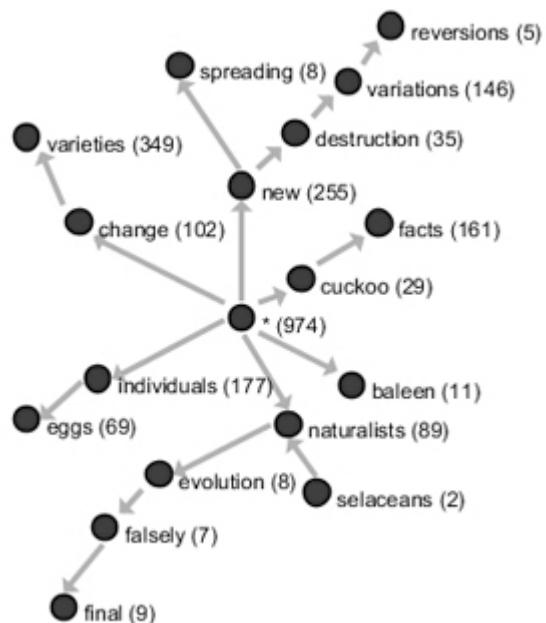


Figure 1: Conceptual analysis of "evolution"

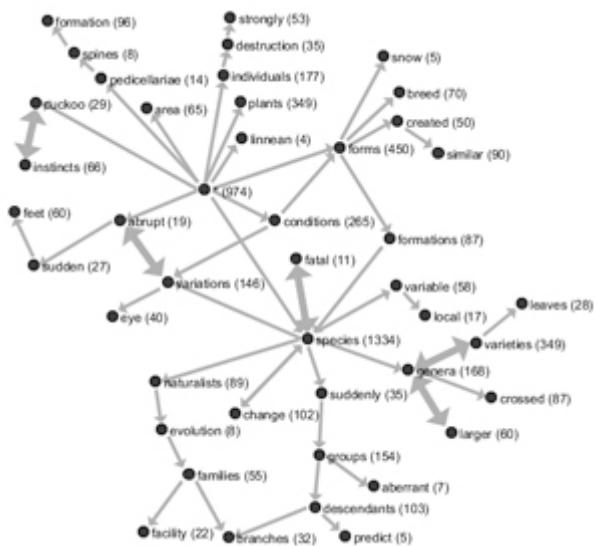


Figure 2: Conceptual analysis of "evolution", "evolve", and "evolved"

In addition to new and unforeseen methodological discoveries, interpretation of both conceptual analyses seems to bring the sixth edition of the *Origin of Species* closer to the contemporary works of the more mature Herbert Spencer, who began to de-emphasize the connection between embryology and the general process of "evolution" and thus contributed to forge the present, strictly specific and most commonly known biological use of the word "evolution".

These results, along with the method that made them possible, are not in any way definitive, and further improvements and modifications of the iterative concordance clustering process are to be expected. Upon completion, this rather new and original approach, while hoping to bring new insights in the understanding of the *Origin of Species*, also aims at underlining the pertinence and usefulness of text mining methods and applications for expert and specialized text reading and analysis, as well as their importance for the future development of philology, hermeneutics, social sciences and humanities in general.

References

- Bowler, P.J.** (2003). *Evolution: the History of an Idea*. Berkeley: University of California Press.
- The Complete Works of Charles Darwin Online*. <http://darwin-online.org.uk> (accessed 25 February 2010).
- Harris, Z.** (1991). *A Theory of Language and Information: A Mathematical Approach*. Oxford: Clarendon Press.

Huxley, T.H. (1894). 'Evolution in Biology'. *Collected Essays, vol II: Darwiniana*. London: Macmillan, pp. 187-226.

MacQueen, J. B. (1967). 'Some Methods for classification and Analysis of Multivariate Observations'. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, pp. 281-297. <http://projecteuclid.org/euclid.bsmsp/1200512992>.

Meunier, J.G., Forest D. and Biskri, I. (2005). 'Classification and Categorization in Computer-Assisted Reading and Text Analysis'. *Handbook of Categorization in Cognitive Science*. Cohen, H. and Lefebvre, C. (ed.). The Hague: Elsevier, pp. 955-978.

Network Workbench Tool. <http://nwb.slis.indiana.edu>.

Richards, R.J. (1992). *The Meaning of Evolution: the Morphological Construction and Ideological Reconstruction of Darwin's Theory*. Chicago: University of Chicago Press.

The TEI's Extramural Journal Project: Exploring New Digital Environments and Defining a New Genre in Academic Publishing

Schlitz, Stephanie A.

sschlitz@bloomu.edu

Bloomsburg University

The Text Encoding Initiative's (TEI) Extramural Journal (EJ) project was conceived early in 2009 when the conveners of the TEI Education Special Interest Group (SIG) proposed, as a matter of urgency, the development of an online publishing suite to address the shortage of TEI educational resources.¹ Following approval by the TEI Board and Council and the receipt of a small SIG grant in support of the project, TEI-EJ advanced into development.

Because TEI-EJ is being researched and developed in an era where “print is no longer the exclusive or the normative medium in which knowledge is produced and/or disseminated” (“A Digital Humanities Manifesto”) and where electronic publication is increasingly common (see Waltham; Maron and Smith; Willett), it is crucial to point out that typologically, TEI-EJ is positioned outside of two disparate points on the web publishing continuum, media-driven journals which are designed primarily for the publication of media-driven content (e.g. *Vectors Journal* <<http://www.vectorsjournal.org/>>, *Southern Spaces* <<http://www.southernspaces.org/>>; also see Toton and Martin) and text-driven journals which are designed to reproduce the print journal model in a web publication (e.g. *Journal of Writing Research* <<http://www.jowr.org>>, *International Journal of Teaching and Learning in Higher Education* <<http://www.isetl.org/ijt1he/>>).

Although TEI-EJ's 'journal' designation is suggestive of a single aim, from the outset, project objectives have been defined as both experimental and extramural, and TEI-EJ has been envisaged not only as a publishing venue but also as a community-driven online forum that offers members of the TEI, whether novice or expert, as well as the broader DH community new educational insights into the TEI. Significantly, the steps being taken to achieve these objectives contribute to a newly emerging body of scholarship which explores the development of new

digital environments and which defines a new genre in academic publishing.

The first stage of the project has been the development of TEI-EJ as a born digital, open access, peer reviewed scholarly journal where communicative modes are bidirectional rather than exclusively unidirectional, articles are media-driven (including video, audio, and image) as well as text-driven, and where the aims of publication extend beyond print journal mimesis to include education and community building.

Given the hybrid nature of the project's goals (publishing and learning community; see Dal Fiore, Koku and Wellman) and the novel design for implementation, TEI-EJ's site infrastructure (see **Fig. 1**) was designed to be extensible, capable of managing text articles (e.g. TEI-XML as well as formats such as .txt and .doc which are converted to TEI-XML), multimedia articles (e.g. video, audio, image), moderated responses to articles, and community-driven communications (e.g. forum and blog) (see **Fig. 2**). This is achieved through Drupal,² a customizable, open source content management system, which facilitates the social media as well as the publishing and educational aspects of the project.³



Fig. 1. *Schlitz TEI-EJ project planning 'mindmap'*



Fig. 2. Screenshot of TEI-EJ website

This paper will introduce the TEI-EJ project, describing the *why* and *how* of the key theoretical, technological and editorial decisions that drove development as we advanced from theory into practice. In doing so, it aims to establish the project as a new model for academic publishing which is designed to harness emerging technologies, to leverage the fact that "Open access is changing the public and scholarly presence of the research article" (Willinsky), to promote learning objectives beside dissemination of scholarship, and to elevate the role of reader/end-user to the position of chief stakeholder.

References

A Digital Humanities Manifesto. 15 Dec. 2008 23 July 2009 <http://dev.cdh.ucla.edu/digitalhumanities/2008/12/15/digital-humanities-manifesto/>.

Dal Fiore, Filippo (2007). 'Communities Versus Networks: The Implications on Innovation and Social Change'. *American Behavioral Scientist.* **(50)7:** 857-866.

Koku, Emmanuel F., Wellman, Barry (2002). 'Scholarly Networks as Learning Communities: The Case of TechNet'. *Designing Virtual Communities in the Service of Learning.* Sasha Barab, Rob Kling (eds.). Cambridge: Cambridge University Press.

Maron, Nancy L., Kirby Smith, K. (2009). 'Current Models of Digital Scholarly Communication: Results of an Investigation Conducted by Ithaka Strategic Services for the Association of Research Libraries'. *The Journal of Electronic Publishing.*

- 12(1).** <http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0012.105>.

TEI: Text Encoding Initiative. <http://www.tei-c.org/index.xml> (accessed 7 July 2009).

Terras, Melissa, Van den Branden, Ron, Vanhouffe, Edward (2009). 'Teaching TEI: The Need for TEI by Example'. *Literary and Linguistic Computing.* **24(3):** 297-306.

Toton, Sarah, Martin, Stacey (2009). 'Teaching and Learning from the U.S. South in Global Contexts: A Case Study of Southern Spaces and Southcomb'. *Digital Humanities Quarterly.* **3(2).** <http://digi talhumanities.org/dhq/vol/3/2/000047.html>.

Waltham, Mary (12 Oct. 2009). 'The Future of Scholarly Journals Publishing'. <http://www.nhali ance.org/bm~doc/hssreport.pdf>.

Willett, Perry (2004). *Electronic Texts: Audiences and Purposes.* 'A Companion to Digital Humanities'. Schreibman, Susan , Siemens, Ray , Unsworth, John (eds.). Oxford: Blackwell. <http://digitalhumaniti es.org/companion/>.

Willinsky, J. (25 Sept. 2009). '9 Flavors of Open Access'. *E-MEDICINE.* **49 (3).** <http://cssp.us/pf/9%20Flavors%20of%20Open%20Access.pdf>.

Notes

1. Two of the notable TEI teaching resources which are available include the Women Writers Project's NEH-funded series of "Advanced Seminars on Scholarly Text Encoding" (see <http://www.wwp.brown.edu/encoding/seminars/neh_advanced.html>) and TEI by Example (see <<http://www.kantl.be/ctb/project/2006/tei-ex.htm>> and Terras et al.).
2. For a good, comparative discussion of content management systems, see "Comparing Open Source Content Management Systems: WordPress, Joomla, Drupal, and Plone," available at <http://idealware.org/comparing_os_cms/>. TEI-XML content is being handled by the Drupal XML Content module, and the journal's publishing workflow is being handled by the Drupal E-Journal module.
3. A preview of the publishing website and a fully functional mock journal issue are being presented at the TEI Members' Meeting in November 2009.

The Specimen Case and the Garden: Preserving Complex Digital Objects, Sustaining Digital Projects

Schlosser, Melanie

schlosser.40@osu.edu

University Libraries, The Ohio State University

Ulman, H. Lewis

ulman.1@osu.edu

Department of English, The Ohio State University

In a recent article entitled "Innkeeper at the Roach Motel", Dorothea Salo worries that focusing exclusively on preservation when designing institutional repositories leads to a situation in which documents are placed into repositories but never come out (Salo 2009). In this conceit, a "live" project gets placed in a repository and "dies" from lack of use. However, when attempting to preserve distributed, dynamic electronic textual editions, a somewhat different metaphor is needed. Like items in a specimen case, "live" digital projects must be "killed" before they are added to a conventional institutional repository such as DSpace. In such applications, they must be removed from the dynamic ecology of their production environments (the garden of our title) and frozen in a snapshot form that is substantially different in appearance and functionality. In our presentation, we will outline our own efforts to construct a preservation and sustainability plan for a multi-format, distributed, dynamic electronic textual edition that involves both the creation of a preservation "specimen" and the careful tending of the edition in the "garden". We will also share the general tools and workflows developed by the project that can help others tackle the same challenges.

Due to their innovative nature and the environments in which they are created, Digital Humanities projects are often dynamic and distributed. In other words, they often exist in multiple parts maintained on distributed hardware (which itself is supported by multiple organizations), and are often compiled for viewing on the fly in response to readers' actions. The typical preservation strategy of frequent offsite backups is inadequate for these projects. The user or manager must be able to find the backups (including all of the constituent parts of a complex project, wherever they reside) and recognize what they are; the backups must be in usable condition; and their contents need to be understandable to the people who want to use them. Moreover, if backup files are to be used in any way similar to their original

use, the files must be (1) compatible with current hardware and software, (2) translated into formats that are compatible with current hardware and software, or (3) used on reconstructed or emulated hardware and software that match the environment in which the project was originally developed.

Digital Humanities projects created by faculty often have the added vulnerability of relying on the creator's university computing accounts. Absent special accommodations, these accounts usually expire on a set schedule once the individual has moved on, taking with them information that often exists nowhere else.

1. Cultures of Preservation

In short, digital materials require a culture of description, preservation, and access every bit as robust as the practices and institutions that allow us to preserve manuscript and print materials. The devil of preservation — whether of print, digital, or other material artifacts — lies in the details of production, use, description, storage, conservation, and access. This holds true whether we are talking about acidic paper disintegrating on library shelves, digital files in obsolete formats, or media spread across computer systems whose links to one another have been broken. Preservation is further complicated by the distinction between preserving physical artifacts (books, manuscripts, floppy disks, flash drives) and preserving the information contained on those media in a useful format.

"The Specimen Case and the Garden: Preserving Complex Digital Objects, Sustaining Digital Projects" focuses on the preservation challenges posed by complex digital humanities projects, which present unique challenges to libraries and repositories charged with accessioning, describing, and preserving the scholarly record. Our work, funded by the U. S. National Endowment for the Humanities, takes a two-pronged approach to the problem, developing technologies for preserving digital *objects* — and the relationships among them — that constitute complex projects, and establishing institutional structures for sustaining digital humanities *projects* after their creators are no longer actively involved in their development and maintenance. Over the course of more than a year, we have interviewed faculty involved in digital humanities projects, library professionals, and information technology professionals; assessed the need for new practices adapted to digital preservation at our institution; and documented the resources and workflows currently available for, or adaptable to, long-term preservation of digital objects. We have also begun to develop tools, institutional structures, and workflows for describing and archiving complex

digital objects, as well as sustaining distributed digital production environments.

2. Preservation and Sustainability Tools and Workflows

Our presentation will outline the problems associated with preserving and sustaining complex digital projects, review the data we collected during our interviews, literature review, and environment scan, and share the tools that we have developed, including the following:

- a lifecycle map of complex digital projects that represents development and preservation milestones as interactions among scholars, library professionals, and IT professionals;
- a visual content manifest for complex digital projects that represents assets, the hardware on which those assets rely, and entities that enable the collaborative work of developing and preserving digital humanities projects;
- a Metadata Encoding and Transmission Standard (METS) profile for creating archival packages of complex digital projects;
- a visual representation of the roles of the scholars, library professionals, and IT professionals on our campus in the long-term preservation of digital humanities projects;
- a proposal for a Digital Humanities Network that could sustain selected distributed digital projects, without requiring that they sacrifice functionality for centralization.

Some of these tools and workflows will be easier to adapt to different projects, institutions, and cultural settings than others: for example, any library system should be able to adapt the METS profile to their needs, while our proposed Digital Humanities Network will serve mostly as a heuristic. Indeed, we hope to initiate a fruitful conversation about how to build cultures of preservation for complex digital projects among scholars, librarians, and IT professionals in a variety of institutional settings.

References

Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2008). 'Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation'. <http://www.citeulike.org/user/AlisonBabeu/article/3801593>.

Cundiff, M. V. (2004). 'An Introduction to the Metadata Encoding and Transmission Standard (METS)'. *Library Hi Tech*. **22.1**: 52-64.

Dobreva, M. (2009). *Digitisation of Special Collections: Mapping, Assessment, Prioritisation, Final Report*. University of Strathclyde: Centre for Digital Library Research (CDLR).

Knight, G. (2009). *SHERPA Digital Preservation 2: Developing Services for Archiving and Preservation in a Distributed Environment, Final Report*. London: JISC; Centre for e-Research, King's College London.

The Library of Congress (2009). <http://www.loc.gov/standards/mets/mets-home.html>

Maron, N. L., Smith, K. K., and Loy, M. (2009). *Sustaining Digital Resources: An on-the-Ground View of Projects Today*. London: JISC: Ithaka Case Studies in Sustainability.

Rieger, O. Y., and Kehoe, B. (DATE). 'Enduring Access to Digitized Books: Organizational and Technical Framework'. *The Fifth International Conference on Preservation of Digital Objects. Joined Up and Working: Tools and Methods for Digital Preservation*. Farqhar, Adams (ed.). London: The British Library.

Robertson, R. J., Mahey, M., and Allinson, J. (2008). *An Ecological Approach to Repository and Service Interactions*. Strathclyde, Scotland: JISC Centre for Educational Technology & Interoperability Standards.

Salo, D. (2008). 'Innkeeper at the Roach Motel'. *Library Trends*. **57**: 2.

Skinner, K., and M. Halbert (eds) (2008). *Strategies for Sustaining Digital Libraries*. Atlanta, GA: Emory University.

Thompson, D. (2008). 'Archiving Web Resources'. *DCC Digital Curation Manual*. Ross, S. and Day, M. (ed.) .

A Tale of Two Cities: Implications of the Similarities and Differences in Collaborative Approaches within the Digital Libraries and Digital Humanities Communities

Siemens, Lynne

siemensl@uvic.ca

Faculty of Business/School of Public
Administration, University of Victoria

Cunningham, Richard

richard.cunningham@acadiau.ca

Acadia Digital Culture Observatory, Acadia
University

Duff, Wendy

wendy.duff@utoronto.ca

Faculty of Information, University of Toronto

Warwick, Claire

c.warwick@ucl.ac.uk

Department of Information Studies, University
College London

Besides drawing on content experts, librarians, archivists, developers, programmers, managers, and others, many emerging digital projects also pull in disciplinary expertise from areas that do not typically work in team environments. To be effective, these teams must find processes – some of which are counter to natural individually-oriented work habits – that support the larger goals and group-oriented work of these digital projects. This paper will explore the similarities and differences in approaches within and between members of the Digital Libraries (DL) and Digital Humanities (DH) communities by formally documenting the nature of collaboration in these teams. The objective is to identify exemplary work patterns and larger models of research collaboration that have the potential to strengthen this positive aspect of these communities even further, while exploring the key differences between them which may limit digital project teams' efforts. Our work is therefore 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.

1. Context

Traditionally, research contributions in the humanities field have been felt to be, and documented to be, predominantly solo efforts by academics involving little direct collaboration with others, a model reinforced through doctoral studies and beyond (See, for example, Cuneo 2003; Newell and Swan 2000). However, DL and DH communities are exceptions to this. Given that the nature of digital projects involves computers and a variety of skills and expertise, collaborations in these fields involve individuals within their institutions and with others nationally and internationally. Such collaboration typically must coordinate efforts between academics, undergraduate and graduate students, research assistants, computer programmers and developers, librarians, and other individuals as well as financial and other resources. Further, as more digital projects explore issues of long term sustainability, academics and librarians are likely to enter into more collaborations to ensure this objective (Kretzschmar Jr. and Potter 2009).

Given this context, some research has been done on the DL and DH (See, for example Liu and Smith 2007; Ruecker and Radzikowska 2008; Siemens 2009) communities as separate entities (See, for example Johnson 2009; Liu, Tseng and Huang 2005; Johnson 2005; Siemens et al. 2009b), but little has been done on the interaction between these two communities when in collaboration. Tensions can exist in academic research teams when the members represent different disciplines and approaches to team work (Birnbaum 1979; Fennel and Sandefur 1983; Hara et al. 2003). Collaborations can be further complicated when some team members have more experience and training in collaboration than other members, a case which may exist with digital projects involving librarians and archivists, who tend to have more experience, and academics, who tend to have less. Ultimately, too little is known about how these teams involving DL and DH members collaborate and the types of support needed to ensure project success.

2. Methods

This paper is part of a larger project examining research teams within the DH and DL communities, led by a team based in Canada and England (For more details, see Siemens et al. 2009a; Siemens et al. 2009b). It draws upon results from interviews and two surveys of the communities exploring the individuals' experiences in digital project teams. The findings include a description of the communities' work patterns and relationships and the identification of supports and research

preparation required to sustain research teams (as per Marshall and Rossman 1999; McCracken 1988). A total of seven individuals were interviewed and another 69 responded to the two surveys.

3. Preliminary Findings

At the time of writing this proposal, final data analysis of the surveys and interviews is being completed. However, some preliminary comparisons between the two communities can be reported.

As a starting point, similarities exist among DL and DH projects. First, digital projects are being accomplished within teams, albeit relatively small ones, as defined by budget and number of individuals involved. Both communities report that the scale and scope of digital projects require individuals with a variety of skills and expertise. Further, these collaborations tend to operate without formal documentation that outline roles, responsibilities, decision making methods, and conflict resolution mechanisms. The survey and interview respondents from both communities report similar benefits and challenges within their collaborations. Finally, these teams rely heavily on email and face-to-face interaction for their project communications.

Some interesting differences between DL- and DH-based teams exist and may influence a digital project team's effectiveness. First, the DL respondents seem to have a greater reliance on email as opposed to face-to-face communications and tend to rate the relative effectiveness of email higher than the DH respondents. Several explanations may be offered for this. According to survey results, DL teams appear more likely to be located within the same institution, which means that casual interpersonal interaction may be more likely to occur between team members than with groups that are geographically dispersed, as many DH teams are. For dispersed teams, meetings need to be more deliberately planned, which may mean a higher consciousness about the importance of this kind of interaction and the necessity to build this into project plans. Also, given that many of the DL teams are within the same organization, team members may be more familiar with each other in advance of a project start, meaning that more communication can be done by email. Less time may need to be spent in formal meetings developing work processes as is the case with those teams whose members may not have worked together on previous projects.

Second, a greater percentage of respondents (42%) within the DH community indicated that they "enjoyed collaboration" than the DL respondents (18%). Comprising of more academics, the DH community tends to undertake more solitary work, and therefore collaboration may be seen as a

welcomed change and may be a deliberate choice that they have made to undertake this type of work. In contrast, team work is more the norm for librarians and archivists, and thus they may feel it is an expected part of their jobs, rather than a choice and welcomed activity. As a result, members of these two communities approach collaboration from two fundamentally different positions, which must be understood from the outset of a digital project in order to reduce challenges and ensure success.

Further, differences in roles and perceived status may complicate collaboration. Often, tensions may exist between service departments, such as libraries and computer support, and the researcher, who is perceived to have higher status (Warwick 2004). These differences in perceived status can complicate work process as those with lower status may have difficulty directing those with perceived higher status (Hagstrom 1964; Ramsay 2008; Newell and Swan 2000).

The benefits to the DL and DH communities will be several. First, the study contributes to an explicit description of these communities' work patterns and inter-relationships. Second, it 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

- Birnbaum, Philip H.** (1979). 'Research Team Composition and Performance'. *Interdisciplinary Research Groups: Their Management and Organization*. Richard T. Barth and Rudy Steck (ed.). Vancouver, British Columbia: International Research Group on Interdisciplinary Programs.
- Cuneo, Carl** (November 2003). 'Interdisciplinary Teams - Let's Make Them Work'. *University Affairs*. 18-21.
- Fennel, Mary, and Gary D. Sandefur** (1983). 'Structural Clarity of Interdisciplinary Teams: A Research Note'. *The Journal of Applied Behavioral Science*. 19.2: 193-202.
- Hagstrom, Warren O.** (1964). 'Traditional and Modern Forms of Scientific Teamwork'. *Administrative Quarterly*. 9: 241-63.
- Hara, Noriko, 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*. 54.10: 952-65.

Johnson, Ian M. (2005). 'In the Middle of Difficulty Lies Opportunity' - Using a Case Study to Identify Critical Success Factors Contributing to the Initiation of International Collaborative Projects'. *Education for Information*. **23. 1/2:** 9-42.

Johnson, Ian M. 'International Collaboration between Schools of Librarianship and Information Studies: Current Issues'. *Asia-Pacific Conference on Library & Information Education & Practice*.

Kretzschmar Jr., William A., and William G. Potter (2009). 'Library Collaboration with Large Digital Humanities Projects'. *Digital Humanities*.

Liu, Jyi-Shane, Mu-Hsi Tseng, and Tze-Kai Huang (2005). 'Building Digital Heritage with Teamwork Empowerment'. *Information Technology & Libraries*. **24.3:** 130-40.

Liu, Yin, and Jeff Smith (2007). 'Aligning the Agendas of Humanities and Computer Science Research: A Risk/Reward Analysis'. *SDH-SEMI*.

Marshall, Catherine, and Gretchen B. Rossman (1999). *Designing Qualitative Research*. Thousand Oaks, CA: SAGE Publications, 3rd edition.

McCracken, Grant (1988). *The Long Interview. Qualitative Research Methods*. Newbury Park, CA: SAGE Publications. V. 13.

Newell, Sue, and Jacky Swan (2000). 'Trust and Inter-Organizational Networking'. *Human Relations*. **53.10:** 1287-328.

Kretzschmar Jr., William A., and William G. Potter (2008). 'Rules of the Order: The Sociology of Large, Multi-Institutional Software Development Projects'. *Digital Humanities*.

Ruecker, Stan, and Milena Radzikowska (2008). 'The Iterative Design of a Project Charter for Interdisciplinary Research'. *DIS*.

Siemens, Lynne (2009). 'It's a Team If You Use "Reply All": An Exploration of Research Teams in Digital Humanities Environments'. *Literary & Linguistic Computing*. **24.2:** 225-33.

Siemens, Lynne, et al. (2009). 'Able to Develop Much Larger and More Ambitious Projects: An Exploration of Digital Projects Teams'. *DigCCurr 2009: Digital Curation: Practice, Promise and Prospects*. Helen R. Tibbo, et al. (ed.). University of North Carolina at Chapel Hill.

Siemens, Lynne, et al. (2009). 'Building Strong E-Book Project Teams: Processes to Maximize Success While Drawing on Essential Academic Disciplinary Expertise'. *BooksOnline '09: 2nd Workshop on Research Advances in Large Digital Book Collections*.

Warwick, Claire. 'No Such Thing as Humanities Computing? An Analytical History of Digital Resource Creation and Computing in the Humanities'. *Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computing*.

Unfolding History with the Help of the GIS Technology: a Scholar-Librarian Quest for Creating Digital Collections

Smith, Natasha

nsmith@email.unc.edu

University of North Carolina at Chapel Hill

Allen, Robert

rallen@email.unc.edu

University of North Carolina at Chapel Hill

Whisnant, Anne

anne_whisnant@unc.edu

University of North Carolina at Chapel Hill

Eckhardt, Kevin

kevineck@email.unc.edu

University of North Carolina at Chapel Hill

Moore, Elise

elimoore@email.unc.edu

University of North Carolina at Chapel Hill

Carolina Digital Library and Archives (CDLA) and Documenting the American South (DocSouth) are a digital library laboratory that creates, develops, and maintains online collections regarding the history of the American South with materials drawn primarily from the outstanding archival holdings of the UNC library. In this presentation, we plan to demonstrate how the close partnership between UNC librarians and faculty forges its path in the frontier of digital humanities. Our experience clearly demonstrates that digital historical scholarship cannot be done on the old model of the scholar laboring alone, “the solitary scholar who churns out articles and books behind the closed door of his office” (see Kenneth Price, 2008). By bringing together faculty and librarians’ expertise, collaborators endeavor to use digital technologies in a variety of innovative ways to collect, organize, and display data and materials that illuminate the temporal and spatial unfolding of historic events. Recent experimental work with GIS helps us to better understand how the use of digital technologies changes the way we do research in humanities and how it facilitates learning in the classroom. Indeed, “GIS, in combination with other branches of scholarship, has the potential to provide a more integrated understanding of history” (see Ian N. Gregory, 2003).

At the same time, the wide array of issues (digitizing and geo-referencing of Sanborn and other historic maps, use of JavaScript mapping APIs, such as Google Maps and the open-source Open Layers, for zooming and hotspot addition, layering and geo-tagging scholarly content) will be presented based on several completed and in progress digital history collections built in close collaboration of UNC librarians working with UNC scholars.

1. “Going to the Show” (<http://www.docsouth.unc.edu/gtts/>) is the first digital archive devoted to the early experience of cinema across an entire state. In a research project, Prof. Allen collaborated with digital publishing experts and special collections librarians at UNC to create an online, interactive digital collection of maps, photos, postcards, newspaper clippings, architectural drawings, city directories and historical commentary that illuminate and reconstruct cultural and social life during the first three decades of the 20th century in North Carolina. Supported by a grant from the N.C. State Library and a National Endowment for the Humanities Digital Humanities Fellowship, “Going to the Show” (GttS) developed the innovative system for layering content on electronically stitched and geo-referenced Sanborn Fire Insurance Maps. Especially in its highly detailed case study of early moviegoing in Wilmington, N.C., GttS demonstrates the extraordinary potential for illuminating community history through the interaction of documentary material and Sanborn maps (see Figure 1).

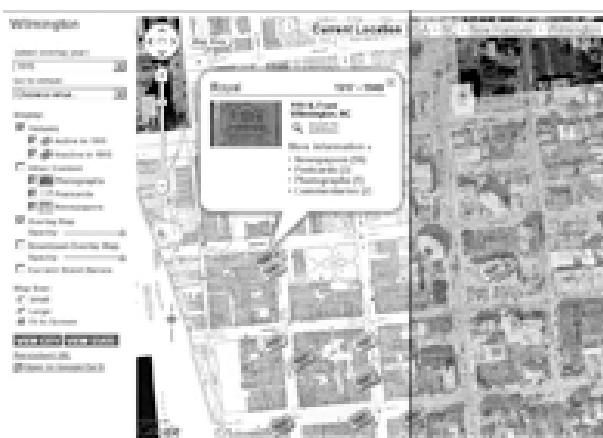


Figure 1. Google Maps API used to present 1915 Sanborn map with layered historic materials to document the moviegoing in North Carolina.

2. Building on the digital history project “Going to the Show”, the project team decided to expand the reach of their expertise by creating a web-based toolkit that will allow libraries, schools, museums, local history societies, and other community organizations to preserve, document, interpret, display, and share the history of their downtowns. Called “*Main Street, Carolina: Recovering the History of Downtown Across North Carolina*,” the toolkit will provide

users with a flexible, user-friendly digital platform on which they can add a wide variety of “local” data: historical and contemporary photographs, postcards, newspaper ads and articles, architectural drawings, historical commentary, family papers, and excerpts from oral history interviews—all keyed to and layered on top of digitized Sanborn Fire Insurance Maps. The toolkit will consist of a PHP-based web application and a JavaScript API. The web application will be compact in size, resource-light, and easy to install on the local organization’s own web server or that of a third-party web-hosting service. It will provide administrative tools for configuring the site, creating place markers, creating simple web pages for content, and customizing the look and feel (see Figure 2). These place markers can then be associated with images, stories, or other content, providing a visual link between the content and related physical locations. The map interface is the focal point of the software and will allow users to explore content associated with specific geographic locations by interacting with place markers, or “pushpins,” overlaid on top of historic maps. Clicking on a place marker’s icon will display an information bubble which can contain text, images, and links to additional content. The user will be able to view and effortlessly pan across entire downtowns as a seamless integration of multiple high-resolution map pages; zoom from a bird-eye view to the smallest cartographic feature; compare successive map iterations showing the same building, block, or neighborhood; and overlay any of these views with contemporary satellite and map images at the same scale. The JavaScript API will allow users to include digitized maps created for other CDLA projects as layers in their own websites or mash-ups which use the Google Maps or Open Layers mapping APIs. For example, a user could embed an historic map in a blog post or add a Sanborn Map as a layer to their existing website which uses Google Maps to show the location of homes that are listed on the *National Register of Historical Places*. MSC is funded by a private funding and an NEH Start-up Grant. Development for this project began in October 2009. We plan to release the toolkit and pilot projects developed in collaboration with external partners in summer 2010, prior to the start of the conference.



Figure 2. “Main Street, Carolina: Recovering the History of Downtown Across North Carolina” tool kit. Administrative form for entering historical documents.

3. “Driving through Time: The Digital Blue Ridge Parkway in North Carolina” will present an innovative visually and spatially based model for illustrating North Carolina’s key role in creating the Parkway, representing the twentieth-century history of a seventeen-county section of the North Carolina mountains, and for understanding crucial elements of the development of the American National Park system. The project will feature historic maps, photographs, postcards, government documents, oral history interviews, and newspaper clippings. Each historic document will be assigned geographic coordinates so that it can be viewed on a map, enabling users to visualize and analyze the impact of the Blue Ridge Parkway on the people and landscape in western North Carolina over both space and time (See Figure 3). Primary sources will be drawn from the collections of the UNC-Chapel Hill University Library, the Blue Ridge Parkway Headquarters, and the North Carolina State Archives. These materials are especially significant in that they document one of North Carolina’s most popular tourist attractions, but also in the way that they help illuminate the way that the Blue Ridge Parkway transformed the communities through which it passed. In addition to the digitized primary sources, the project will include scholarly analyses of aspects of the development of the Blue Ridge Parkway.

A geospatial format is uniquely appropriate for considering the history of the Parkway and its region. As a narrow park corridor pushed through a long-populated southern Appalachian landscape, the Parkway rearranged spaces, repurposed lands, reorganized travel routes, and opened and closed economic opportunities through control of road routing, access, and use. The social conflicts it engendered, therefore, frequently entailed spatial components – should the road go here, or there; should it take more or less land; should this or

that property be favored with direct Parkway access (or not)? Understanding these aspects of Parkway history without reference to spatial relationships on the land is challenging, as the project's scholarly adviser, Dr. Anne Mitchell Whisnant, recognized when publishing her 2006 book, *Super-Scenic Motorway: A Blue Ridge Parkway History* (UNC Press). Her experience, both in writing the book and in delivering numerous public presentations since its appearance, is that narrative alone cannot provide the public with the tools to comprehend past controversies or present land protection challenges. Using digital and geospatial technologies to open a new window on the history of the Parkway and its region is especially timely considering the approach of the Parkway's 75th anniversary in 2010 and the National Park Service's 100th anniversary in 2016.

The collaboration between the library and Dr. Whisnant has been enhanced by Whisnant's engagement in related field of "public history," which is history practiced outside the walls of academia, with and for public audiences. Based in academia but designed for public benefit, "Driving through Time" has offered an exceptional opportunity for involving undergraduate and graduate students in Dr. Whisnant's Introduction to Public History class in its creation. As a scholarly project being built through the expertise of a large team (as nearly all public history undertakings are), "Driving through Time" has been an ideal space for students to gain hands-on experience in doing public history collaboratively, in real-time, with their instructor. Students are doing original primary source research in the university's special collections, identifying materials for inclusion in the online exhibit, developing their own historical narratives, working with new tools such as wikis and databases, and contributing to the creation of metadata. Because the instructor has not predetermined the final outcome, furthermore, students are being given ownership over both their process and their final products and practicing navigating the unexpected twists, turns, delights, and disappointments that historical research always entails.



Figure 3. "Driving through Time: The Digital Blue Ridge Parkway in North Carolina"

References

- Knowles, Anne Kelly** (2008). *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*. Redlands, Calif.: ESRI Press.
- Whisnant, Anne Mitchell** (2006). *Super-Scenic Motorway: A Blue Ridge Parkway History*. UNC Press.
- Gregory, Ian N.** (2003). *A place in History: A Guide to Using GIS in Historical Research*. 2nd Edition. <http://www.ccsr.ac.uk/methods/publications/ig-gis.pdf>.
- Price, Kenneth** (2008). 'Electronic Scholarly Editions'. *A Companion to Digital Literary Studies*. Schreibman, Susan, Siemens, Ray (eds.). Oxford: Blackwell.

WW1 and WW2 on a Specialist E-forum. Applying Corpus Tools to the Study of Evaluative Language

Sokół, Małgorzata

msokol@autograph.pl

Szczecin University, Poland

The 70th anniversary of WW2 generated a new discussions about Poles' attitudes to the country's important historical events. The general aim of my present study is to investigate how contemporary Poles' attitudes to the two world wars are expressed in electronic discourse, using the example of a specialist Internet forum. The research data comes from an active Polish forum devoted to the history of the two world wars: *Your Forum About the World Wars* retrieved from <http://www.dws.org.pl>. The quantitative-qualitative analysis of the compiled corpus is conducted by means of the UAM Corpus Tool, a multiple-level environment for annotation and analysis of text corpora (<http://www.wagsoft.com/CorpusTool/>).

The forum under study provides a challenging research context for the study of evaluative language in the electronic medium by means of corpus tools. The forum gathers both professionals and amateurs interested in an academic debate on the issues related to their interests in the history of the two world wars. Interaction on the forum follows the rules of scholarly exchange, where careful use of language and factual expression are encouraged, whereas emotionality is rather disfavoured. This makes textual realizations of evaluation less overt: thus, how attitudinal language is transmitted through the text can be studied effectively by means of corpus tools.

The theoretical basis of my study is a functional approach to the analysis of evaluative language, where evaluation encompasses both attitudinal and affective aspects of language use (e.g. Hunston and Thompson, 1999). For the corpus-based study of evaluative language, I adopt Martin and White's Appraisal Framework (Martin and White, 2005), selecting their systems of Attitude and Engagement. More specifically, I analyse linguistic realisations of affect (within the system of Attitude) and modality and evidentiality (within the system of Engagement) as elements of interactional aspects of language use. In this way I aim to investigate evaluation as 1) expression of both individual and communal value

systems of language users, and 2) expression of the speaker-audience relations (especially with the focus on the rhetorical effects of evaluation and its role in the social construction of knowledge). With the present study I also hope to contribute to the previous research that shows how text analyses within the framework of systemic-functional linguistics can profit from the use of corpus-linguistic methods (Bednarek, 2008; Thompson and Hunston, 2006), at the same time attempting to investigate how the combination of SFL and CL works for the Polish language.

The preliminary results of the study in progress prove the presence of evaluative language across the corpus, with the prevalence of engagement over attitude. This demonstrates that the members of the forum under study are primarily involved in objective and to-the-point discussions on the issues related to their interests. Evaluative language is mainly used as a persuasive tool to enhance the power of knowledge claims, and to manage interpersonal relations. As regards the general attitudes of the forum members, the study proves that they are actively engaged in constructing and preserving the collective memory of the two world wars. The wars are discussed with professional commitment and without overly sentimental pathos, which makes the forum and history of the two world wars attractive also for young people.

The application of the UAM Corpus Tool in the analyses of evaluative language proves advantageous first of all thanks to the tool's possibility of multi-level annotation and cross-classification of features. These features allow me to study the distribution of evaluative language relative to the sub-generic status of postings (i.e. whether a posting is identified as a voice in a discussion, initiation of a discussion, request, query, etc.). In addition, the tool allows to manage the problem of evaluation "hidden" behind stretches of discourse, as well as to analyse the distribution patterns of evaluative lexis. Finally, as the UAM Corpus Tool allows annotation of both text and images, it can be used for the analysis of multimodal genres, an example of which is an e-forum. The UAM Corpus Tool can be useful and flexible in corpus-based analyses of evaluative language, however its application invariably demonstrates the importance of human choice in working with corpora in discourse analysis (cf. also Baker, 2006).

References

- Baker, P. (2006). *Using Corpora in Discourse Analysis*. London - New York: Continuum.

Bednarek, M. (2008). *Emotion Talk Across Corpora*. Hounds Mills - New York: Palgrave Macmillan.

Hunston, S., Thompson, G. (eds.) (1999). *Evaluation in Text. Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press.

Martin, J. R., White, P. R. R. (2005). *The Language of Evaluation. Appraisal in English*. Hounds Mills - New York: Palgrave Macmillan.

Thompson, G., Hunston, S. (eds.) (2006). *System and Corpus. Exploring Connections*. London: Equinox.

Visualization and Analysis of Visiting Styles in 3D Virtual Museums

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

A virtual museum is a cyberspace in persistent virtual worlds, such as Second Life, for displaying digitalized heritage documents. Urban et. al. (2007) reported that over 150 sites in Second Life (SL) were developed for education and museum activities. Virtual museums in SL offer visitors opportunities to engage in opening art exhibitions, discuss with specialists, and enjoy exploring collections of the wide range of artifacts. Those artifacts displayed in the virtual museums vary from 3D documents of the world heritages to fictional creations (Rothfarb, R. and Doherty, P., 2007).

This paper aims at visualization and analysis of visitor behaviors in 3D virtual museums. Without loss of generality, we focused on a museum in Second Life, named Ritsumeikan Digital Archive Pavilion (RDAP) as shown in Figures 1-2. The museum was used in this paper for developing a prototype of our visualization and analysis tool. Efficient visualization of the user movement is very useful for analyzing his/her behaviors, in an implicit manner, in order to extract the disclosed information of individuals in the cyberspace.

Applications of the proposed visualization method include the following:

1. The curators can design the exhibit space based on the majority of visitors as illustrated in Section 3.
2. The storytelling of an individual visitor can be expressed as a sequence of screenshots capturing the most favored exhibits (as described in Fujita, H. et al., 2008).
3. A guide system can be applied so as to achieve a satisfactory museum tour as introduced by Sookhanaphibarn and Thawonmas.



Figure 1. Kimono exhibition in Ritsumeikan
Digital Archive Pavilion (RDAP)

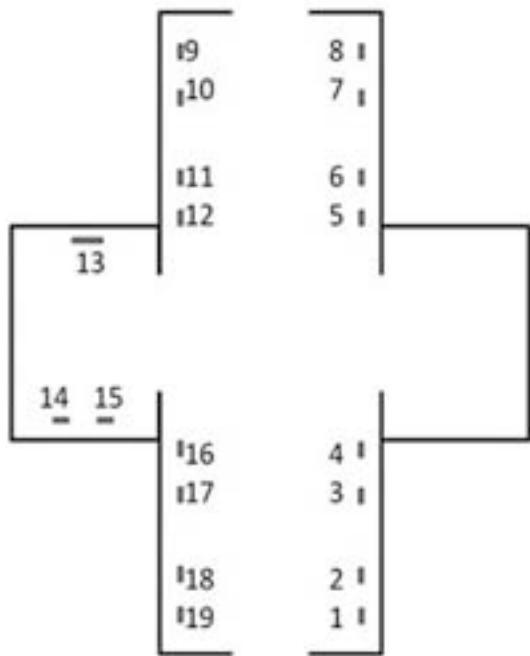


Figure 2. Floor plan of RDAP with the locations of 19 Kimono objects denoted by a small solid square

1. Visualization and analysis of visiting patterns

To validate our visualization approach, 36 avatars' movements in RDAP were synthesized for obtaining four visiting styles. These styles, based on the metaphor of animal behavior, are ant, fish, grasshopper, and butterfly styles as follows (Veron, E. & Levasseur, M., 1983; Chittaro, L., 2004):

1. The ant visitors spend quite a long time to observe all exhibits and walk close to exhibits, but avoid empty spaces.
2. The fish visitors prefer to move to and stop over at empty spaces, but avoid areas near exhibits.
3. The grasshopper visitors spend a long time to see selected exhibits, but ignore the rest of exhibits.
4. The butterfly visitors observe almost exhibits, but spend varied times to observe each exhibit.

2. Local Visualization

Tracing the user movement in Second Life was achieved by using the provided Linden Second Life script, named sensor function. The sensor function detects and reports the user position (x,y) within the particular range. It repeats every particular time interval. In this paper, the considered data consist of the three dimensional positions of an individual visitor and their corresponding time spent.

Figure 3 shows four visiting paths in RDAP. The visualization consists of line segments and white dots. A colored line segment is a part of the avatar movement. A white dot represents the location of a Kimono object.

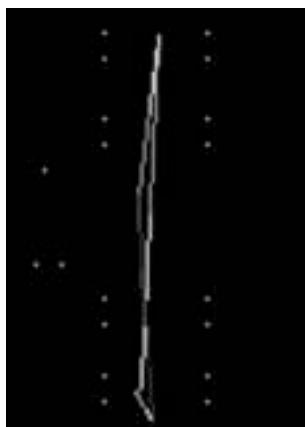
The avatar's path is displayed with the spectrum colors containing red, orange, yellow, green, cyan, blue, and violet. A path is in the form of connecting segments from red to violet. The session is equally divided in time into 7 periods in the ascending order from red to violet. The color of a particular segment indicates the passage of time.

The length of a segment inversely represents the time spent. For example, the avatar as shown in Figure 3 (a) spent the longest time in the first period recognizable from the shortest red segment. The avatar started moving faster during the last two periods as denoted by longer blue and violet. In our visualization, the shorter a segment is, the longer an avatar spends time in that particular area.

Absent colors show that a longer time was spent than one period. For example, no red and orange segments are shown in Figure 3 (b) because the avatar spent time from the first period to the third near the entrance denoted by the yellow area. If an avatar spends too much time at a particular position exceeding the period length, then the color corresponding to that period will be skipped.



(a) Ant visiting style



(b) Fish visiting style



(c) Grasshopper visiting style



(d) Butterfly visiting style

Figure 3: Visualization of four visiting styles in RDAP. These color figures will be provided in URL below. <http://www.ice.ci.ritsumei.ac.jp/~kingkarn/>

3. Analysis

Our visualization approach can describe the aforementioned visiting styles. Ant, Fish, Grasshopper, and Butterfly visiting styles are displayed on totally different vivid graphic graphs. Hence, our tool is useful for distinguishing the visiting types.

The Ant visiting style is shown in Figure 3 (a). The avatar's path was along the white dots. It means that the visitor walked close to Kimono objects in order to look at them in detail. The path contained its segments of nearly equal length, indicating that the visitor spent his/her time with the exhibits almost equally.

The Fish visiting style is shown in Figure 3 (b). The avatar's path was limited to the empty space between two exhibit rows. It means that this visitor preferred to stroll to take the atmosphere of the gallery. Most segments far from the white dots depict that the avatar did not pay attention to the Japanese art imposed on Kimonos. The Grasshopper visiting style is shown in Figure 3 (c). The avatar's path was drawn as a triangle polygon having the smaller area than that of the Ant style. Its segments represent a kind of diagonally walking across the gallery to the interesting exhibits.

The Butterfly visiting style is shown in Figure 3 (d). The avatar's path was the longest path of the four styles. Its segments show plenty of diagonally walking across the gallery to most exhibits. The diagonal segment does not imply his/her preference, but this unorganized visit does not follow a well structured sequence like the Ant.

4. Global Visualization

The global visualization of all visitors in each category is displayed in Figure 4. The methodology of this global visualization consists of

1. visiting style identification,
2. trace accumulation of all users belonging to the same visiting style, and
3. contrast enhancement in order to highlight the most popular route.

Using the synthesized data of 36 visitors, the global visualization of each visiting style can guide curators to rotate the museum items and arrange the sequence of items. Table 1 summarizes the interesting or skipped items associated with the visiting styles. An interesting item (I) can be determined if its observation area is darkened; otherwise, the item (S) is considerably ignored. The item numbers are those assigned in Figure 2.

Curators can design an efficient exhibition based on the majority of visitors. It is assumed that our museums consist of four rooms each of which the majority of visitors are ant, fish, grasshopper and butterfly, respectively.

1. The ant room: the 13th item is possibly not related to others; therefore, a new one should be substituted.

2. The fish room: the visitors prefer to pass slowly through the room. Therefore, all exhibits should be placed along both sides of the main path. The exit should be on either side of the entrance to prevent congestion.
3. The grasshopper room: half of the items are possibly not visitor attraction for busy people; on the other hand, they are perhaps varying and unrelated. Therefore, the curators should be re-design the exhibition room; in addition, the skipped items should be replaced with others more related to those visited.
4. The butterfly room: the visit routes should be unorganized; therefore, the sequence of exhibits should be rearranged and some skipped items should be replaced/removed, accordingly.



(a) Common paths of ant visitors



(b) Common paths of fish visitors



(c) Common paths of grasshopper visitors



(d) Common paths of butterfly visitors

Figure 4. Visualization of common paths based on four visiting styles, each of a group of seven visitors. Kimono exhibits and the user's paths are denoted by red dots and blue lines. The color intensity indicates the frequency of visits. The darker blue the visualization shows, the more visitors spend time at that particular area.

Items	Ant	Grasshopper	Butterfly
1	I	S	I
2	I	S	I
3	I	I	I
4	I	I	I
5	I	I	I
6	I	S	I
7	I	I	I
8	I	S	I
9	I	S	I
10	I	I	S
11	I	S	I
12	I	I	I
13	S	I	S
14	I	S	I
15	I	S	I
16	I	I	I
17	I	I	I
18	I	S	S
19	I	I	I

Table 1. Example of visual analytics from Figure 4 showing the interesting and skipped items in RDAP are as denoted by "I" and "S", respectively

References

- Urban, R., Twidale, M.B. and Marty, P.F.** (2007). 'Second Life for Museums and Archeological Modeling'. *Digital Humanities 2007, Conference Abstracts Book*. <http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=254> (accessed 15 Nov 2009).
- Rothfarb, R. and Doherty, P.** (2007). 'Creating Museum Content and Community in Second Life'. *Museums and the Web 2007: Proceedings*. J. Trant and D. Bearman (ed.). Toronto: Archives & Museum Informatics. <http://www.archimuse.com/mw2007/papers/rothfarb/rothfarb.html> (accessed 3 August 2009).
- Fujita, H. and Arikawa, M.** (2008). 'Animation of Mapped Photo Collections for Storytelling'. *IEICE Trans. INF & SYST. Vol. E91-D, No. 6*: 1681-1692.
- Sookhanaphibarn, K. and Thawonmas, R.** (2009). 'A Movement Data Analysis and Synthesis Tool for Museum Visitors' Behaviors'. *Lecture Notes in Computer Science. Subseries of Information Systems and Applications, incl. Internet/Web, and HCI. Vol. 5879*: 144-154.
- Veron, E. & Levasseur, M** (1983). *Ethnographie de l'exposition. L'espace, le corps et le sens*. Paris: Bibliothque publique d'Information. Centre Georges Pompidou.

Chittaro, L. and Leronutti, L. (2004). 'A Visual Tool for Tracing Users' Behavior in Virtual Environments'. *Proceedings of the Working Conference on Advanced Visual Interfaces, AVI'04 ACM*. Pp. 40-47.

Two representations of the semantics of TEI Lite

Sperberg-McQueen, C. M.

cmsmcq@blackmesatech.com
Black Mesa Technologies LLC, USA

Marcoux, Yves

yves.marcoux@umontreal.ca
Université de Montréal, Canada

Huitfeldt, Claus

Claus.Huitfeldt@uib.no
Department of Philosophy, University of Bergen

Markup languages based on SGML and XML provide reasonably fine control over the syntax of markup used in documents. Schema languages (DTDs, Relax NG, XSD, etc.) provide mature, well understood mechanisms for specifying markup syntax which support validation, syntax-directed editing, and in some cases query optimization. We possess a much poorer set of tools for specifying the *meaning* of the markup in a vocabulary, and virtually no tools which could systematically exploit any semantic specification. Some observers claim, indeed, that XML and SGML are “just syntax”, and that SGML/XML markup has no systematic semantics at all. Drawing on earlier work (Marcoux et al., 2009), this paper presents two alternative and complementary approaches to the formal representation of the semantics of TEI Lite: *Intertextual semantics* (IS) and *Formal tag-set descriptions* (FTSD).

RDF and Topic Maps may appear to address this problem (they are after all specifications for expressing “semantic relations,” and they both have XML transfer syntaxes), but in reality their focus is on generic semantics — propositions about the real world — and not the semantics of markup languages.

In practice, the semantics of markup is most of the time specified only through human-readable documentation. Most existing colloquial markup languages are documented in prose, sometimes systematically and in detail, sometimes very sketchily. Often, written documentation is supplemented or replaced in practice by executable code: users will understand a given vocabulary (e.g., HTML, RSS, or the Atom syndication format) in terms of the behavior of software which supports or uses that vocabulary; the documentation for Docbook elevates this almost to a principle, consistently speaking not of the meaning of particular constructs, but of the “processing expectations” licensed by those constructs.

Yet a formal description of the semantics of a markup language can bring several benefits. One of them is the ability to develop provably correct mappings (conversions, translations) from one markup language to another. A second one is the possibility of automatically deriving facts from documents, and feeding them into various inferencing or reasoning systems. A third one is the possibility of automatically computing the semantics of part or whole of a document and presenting it to humans in an appropriate form to make the meaning of the document (or passage) precise and explicit.

There have been a few proposals for formal approaches to the specification of markup semantics. Two of them are *Intertextual Semantic Specifications*, and *Formal Tagset Descriptions*.

Intertextual semantics (IS) (Marcoux, 2006; Marcoux & Rizkallah, 2009) is a proposal to describe the meaning of markup constructs in natural language, by supplying an IS specification (ISS), which consists in a pre-text (or text-before) and a post-text (or text-after) for each element type in the vocabulary. When the vocabulary is used correctly, the contents of each element combine with the pre- and post-text to form a coherent natural-language text representing, to the desired level of detail, the information conveyed by the document. Although based on natural language, IS differs from the usual prose-documentation approach by the fact that the meaning of a construct is dynamically assembled and can be read sequentially, without the need to go back and forth between the documentation and the actual document.

Formal tag-set descriptions (FTSD) (Sperberg-McQueen et al., 2000) (Sperberg-McQueen & Miller, 2004) attempt to capture the meaning of markup constructs by means of “skeleton sentences”: expressions in an arbitrary notation into which values from the document are inserted at locations indicated by blanks. FTSDs can, like ISSs, formulate the skeleton sentences in natural language prose. In that case, the main difference between FTSD and ISS is that an IS specification for an element is equivalent to a skeleton sentence with a single blank, to be filled in with the content of the element. In the general case, skeleton sentences in an FTSD can have multiple blanks, to be filled in with data selected from arbitrary locations in the document (Marcoux et al., 2009). It is more usual, however, for FTSDs to formulate their skeleton sentences in some logic notation: e.g., first-order predicate calculus or some subset of it.

Three other approaches, though not directly aimed at specifying markup semantics, use RDF to express document structure or *some* document semantics, and could probably be adapted or extended to serve

as markup semantics specification formalisms. They are *RDF Textual Encoding Framework* (RDFTef) (Tummarello et al., 2005) (Tummarello et al., 2006), EARMARK (*Extreme Annotational RDF Markup*) (Di Iorio et al., 2009), and GRDDL (*Gleaning Resource Descriptions from Dialects of Languages*) (Connolly, 2007).

RDFTef and EARMARK both use RDF to represent complex text encoding. One of their key features is the ability to deal with non-hierarchical, overlapping structures. GRDDL is a method for trying to make parts of the meaning of documents explicit by means of an XSLT translation which transforms the document in question into a set of RDF triples. GRDDL is typically thought of as a method of extracting meaning from the markup and/or content in a particular document or set of documents, rather than as a method of specifying the meaning of a vocabulary; it is often deployed for HTML documents, where the information of most immediate concern is not the semantics of the HTML vocabulary in general, but the implications of the particular conventions used in a single document. However, there is no reason in principle that GRDDL could not be used to specify the meaning of a markup vocabulary apart from any additional conventions adopted in the use of that vocabulary by a given project or in a given document.

If proposals for formal semantics of markup are scarce, their application to colloquial markup vocabularies are even scarcer. Most examples found in the literature are toy examples. A larger-scale implementation of RDFTef for a subset of the TEI has been realized by Kepler (Kepler, 2005). However, as far as we know, no complete formal semantics has ever been defined for a real-life and commonly used colloquial vocabulary. This paper reports on experiments in applying ISSs and FTSDs to an existing and widely-used colloquial markup vocabulary: TEI Lite.

Developing an ISS and an FTSD in parallel for the same vocabulary is interesting for at least two reasons. First, it is an opportunity to verify the intuition expressed in Marcoux et al., 2009 that working out ISSs and FTSDs involves much the same type of intellectual effort. Second, it can give insight into the relative merits and challenges of natural-language vs logic-based approaches to semantics specification.

The full paper will focus on the technical and substantive challenges encountered along the way and will describe the solutions adopted.

An example of a challenge is the fact that TEI Lite documents can be either autonomous or transcriptions of existing exemplars. Both cases are treated with the same markup vocabulary, but

ultimately, the meaning of the markup is quite different: in one case, it licences inferences about the marked-up document itself, while in the other, it licences inferences about the exemplar. The work reported in Sperberg-McQueen et al., 2009 on the formal nature of transcription is useful here to decide how to represent statements about the exemplar, when it exists. However, the problems of determining whether any particular document is a transcription or not, and of putting that fact into action in the generation of the semantics remain. One possible solution is to consider as external knowledge the fact that the document is a transcription. In the FTSD case, that external knowledge would be represented as a formal statement that could then trigger inferences about an exemplar. In the ISS case, it would show up as a preamble in the pre-text of the document element. Another solution is to consider the transcription and autonomous cases as two different application contexts of the vocabulary, and define two different specifications. The benefits and disadvantages of the two solutions will be discussed.

Follow-on work will include developing a GRDDL specification of TEI Lite, and comparing it to the ISS and FTSD. It will also include the elaboration of tools to read TEI Lite-encoded documents and generate from them either a prose representation of the meaning of the markup (from the ISS) or a set of sentences in a formal symbolic logic (from the FTSD). We also expect to induce a formal ontology of the basic concepts appealed to by the three formalisms and attempt to make explicit some of the essential relations among the concepts in the ontology: What kinds of things exist in the world described by TEI Lite markup? How are they related to each other?

References

- Connolly, Dan.** *Gleaning Resource Descriptions from Dialects of Languages (GRDDL)*. <http://www.w3.org/TR/grddl/>.
- Di Iorio, A., Peroni, S., Vitali, F.** (2009). 'Towards markup support for full GODDAGs and beyond: the EARMARK approach'. *Balisage: The Markup Conference*. Montréal, Canada, August 2009. Balisage Series on Markup Technologies. 3 vols. <http://doi:10.4242/BalisageVol3.Peroni01>.
- Kepler, F. N.** *RDF Textual Encoding Framework*. <http://sourceforge.net/projects/rdftef/>.
- Marcoux, Y.** (2006). 'A natural-language approach to modeling: Why is some XML so difficult to write?'. *Extreme Markup Languages*. Montréal, Canada, August 2006. <http://conferences.idealliance.org/ex>

treme/html/2006/Marcoux01/EML2006Marcoux01.htm.

Marcoux, Y., Rizkallah, É. (2009). 'Intertextual semantics: A semantics for information design'. *Journal of the American Society for Information Science & Technology*. Volume 60, Issue 9: 1895-1906. <http://doi:10.1002/asi.21134>.

Marcoux, Y., Sperberg-McQueen, C. M., Huitfeldt, C. (2009). 'Formal and informal meaning from documents through skeleton sentences: Complementing formal tag-set descriptions with intertextual semantics and vice-versa'. *Balisage: The Markup Conference*. Montréal, Canada, August 2009. <http://doi:10.4242/BalisageVol3.Sperberg-McQueen01>.

Sperberg-McQueen, C. M., Huitfeldt, C., Marcoux, Y. (2009). 'What is transcription? (part 2) (abstract)'. *Digital Humanities 2009 Conference Abstracts, June 2009*. Pp. 257-260. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferenceproceedings_final.pdf.

Sperberg-McQueen, C. M., Huitfeldt, C., Renear, A. (2000). 'Meaning and Interpretation of Markup: Not as Simple as You Think'. *Extreme Markup Languages 2000*. Montréal, Canada, August 2000.

Sperberg-McQueen, C. M., Miller, E. (2004). 'On mapping from colloquial XML to RDF using XSLT'. *Extreme Markup Languages 2004*. Montréal, Canada, August 2004. http://conferences.idealliance.org/extreme/html/2004/Sperberg-McQueen01/EML2004_Sperberg-McQueen01.html.

Tummarello, G., Morbidoni, C., Kepler, F., Piazza, F., Puliti, P. (2006). 'A novel Textual Encoding paradigm based on Semantic Web tools and semantics'. *5th edition of the International Conference on Language Resources and Evaluation*. Genoa, Italy, May 2006. <http://www.sdjtsi/bib/lrec06/pdf/225.pdf.pdf>.

Tummarello, G., Morbidoni, C., Pierazzo, E. (2005). 'Toward Textual Encoding Based on RDF'. *9th ICCC International Conference on Electronic Publishing*. Leuven-Heverlee, Belgium, June 2005. <http://elpub.scix.net/data/works/att/206elpub2005.content.pdf>.

Thinking Archivally: Search and Metadata as Building Blocks for a New Digital Historiography

Sternfeld, Joshua

jsternfeld@neh.gov

National Endowment for the Humanities, USA

"Records are no longer fixed, but dynamic. The record is no longer a passive object, a 'record' of evidence, but an active agent playing an on-going role in lives of individuals, organizations, and society."¹

Advances in digital representation and preservation have ushered in new perspectives for defining the record. Terry Cook, speaking on behalf of a growing cohort interested in reshaping the disciplinary boundaries of archival studies, argues that records no longer possess the aura of absolute authority that they once held. Practitioners and theorists working within a postmodernist framework have broken open the once sacred bond between the historical craft and the archive, in the process challenging notions of evidence, truth, and narrative.

One of the great beneficiaries and active participants of this re-evaluation of the archive and the record, of course, is digital history. Digital tools, from Omeka to ArcGIS, have empowered a growing community of professional and amateur historians, museums, and libraries to provide unprecedented access to collections of primary materials and historical data. Such tools also demonstrate the ease with which archival and historical practices have come into contact with one another, thereby disrupting conventional understanding of the record in both fields.

In this paper, I will address the cross-disciplinary relationship between history and archival theory as one component in the broader development of a much-needed digital historiography.² I will argue that principles of archival theory and historiography together may guide the evaluation of digital and new media historical representations, especially with regards to the contextualization of historical evidence. Whether considering an online archive, database, or GIS visualization, the aggregation of large data sets necessitates proper archival management of the data. Besides enhancing the long-term sustainability and preservation of the representation – itself a worthy and often overlooked objective –, the application of archival standards to data collection, organization, and

presentation influences the type and quality of conclusions users may generate. Within this complex association among history, archival theory, and digital technology this paper will examine two interrelated "building blocks" – search and metadata – that work hand-in-hand to form the foundation for a sound digital historical representation. How a user queries a representation, even one that is non-textual, governs the quality of historical knowledge at the user's disposal, whereas a representation's content metadata governs the conclusions the user may draw with that knowledge.

1. A Call for Digital Historiography

New techniques to query, sort, catalogue, and visualize historical data have brought renewed interest to understanding the past on every scale, from the personal to the global. As scholars and teachers, we have encouraged digital exploration, whether in gathering local data with the support of a historical society or archive, or repurposing historical materials through museum installations, websites, documentaries, and multimedia mashups.

Despite promising possibilities in historical computing, the emergence of digital history has also created a distinct fissure in the wider field of history. Practicing digital history challenges methodological preconceptions. Conducting search queries across vast digital collections seems antithetical to visiting an archive. Similarly, navigating through a three-dimensional environment enhances interactivity and engagement with the historical representation, and in the process confronts, or at times abandons altogether, the core activities of reading and writing historical texts. In short, history in the digital age has upended notions of representation, context, inquiry, narrative, linearity, temporal and spatial orientation, and experience.³

This undeniable shift in the landscape demands that we harness the potential of digital history while not altogether abandoning established theoretical and methodological practices. A rush to embrace new digital modes of doing history, unfortunately, has overwhelmed a parallel critical examination of changes to these fundamentals. The same techniques and technologies that are laudably tearing down institutional barriers, challenging entrenched theories, and introducing new voices and democratic perspectives, can also advance specious information and theories; distort or obscure the historical record; or worse – eliminate it altogether. The role of the historian, therefore, has shifted from that of exclusive authority to the equally critical role of mediator of historical knowledge. If active participation and exploration have become the benchmarks of digital historical representations, then the (digital) historian

must ensure that the manner of user participation is conducted equitably and responsibly insofar as the knowledge produced through the representation is predicated upon rigorous logic and concentrated historical data.

What principles should a new digital historiography advocate and why is its cultivation imperative? A working digital historiography will enable critical engagement with digital and new media representations, a challenging endeavor considering the spectrum of possible forms that a representation may take. We may justifiably question whether an online collection, for example, shares traits with a GIS-based visualization. While each representational genre warrants a unique set of evaluative criteria, commonalities across formats and historical content do exist and warrant further attention. We may begin with the notion that all representations possess some form of a user interface. Interrogating the user interface can lead one to assess the transparency with which the representation has selected and organized its content. We may also ask whether its formal design complements and provides sufficient access to the content. With a scholarly text, answers to such questions are readily apparent by poring over indicators such as footnotes, bibliography, and the table of contents. Many digital representations, however, collate information within multi-dimensional, non-linear structures, thereby subverting or eliminating such identifiable cues. As Edward Ayers remarks, "We cannot judge a Web site by its cover – or its heft, its publisher's imprint, or the blurbs it wears."⁴

2. The Building Blocks of Digital Historiography: Search and Metadata

In developing a set of evaluative criteria, we must consider the association between a representation's form and content, which together comprise the representation's overall historical argument. While there are numerous components worthy of consideration, two in particular – search and metadata – determine to a large extent how a representation organizes its historical information. Without a robust search engine the user cannot access historical data; similarly, without quality metadata, a strong search engine is rendered ineffective. While this may seem self-evident, the integration of search and metadata in a representation runs much deeper; it affects, and is affected by, nearly every aspect of the representation, including its interface, aesthetic, design, structure, and functionality. Search and metadata together govern the transformative process by which historical information becomes historical evidence.

This paper will use examples of current digital collections and visualizations to illustrate how search and metadata contributes to the overall value of the representation. I will argue that an assessment of these two building blocks, when considered from both an historical and archival perspective, can shed light on the argument put forth by the representation. In the case of an online collection, for example, the creator must weigh the benefits of generating metadata according to standardized thesauri, scholarly input, or folksonomy. These very different approaches, if applied to the same archival collection, would not only influence the type of audience that may use the archive, but also steer users towards divergent search results, which could ultimately determine how the content is recombined.⁵

A reconstruction of an historic building, meanwhile, invites a "search" process of a different sort. Searching occurs while the user navigates through the environment. Is the user invited to discover new sightlines or gauge the distance between structures? If so, does the user have access to previous theories with which to compare a new finding? Even small questions, such as why a virtual archway was set at eight feet instead of six when there may be inconclusive evidence for both, can unearth rich discoveries. The reconstruction thus must make architectural or GIS metadata discoverable, to the extent that this is feasible, in order to foster further investigation.

It is critical that we do not lose sight of the underlying question that should guide the creation and evaluation of all digital historical representations: does the representation invite the user to conduct humanistic inquiry? What are the historical problems encompassed by the representation, and does the evidence compel the user towards addressing those questions or asking new ones? The more the user is made aware of a representation's construction, the greater the potential for productive engagement. Search and metadata thus function as the bridge linking a representation's formal structure and content. Evaluating these two areas along archival and historiographical lines can lead to an assessment of its trustworthiness as a source for generating historical knowledge. In other words, interrogating a representation's search and metadata provides a window to explore a representation's construction of historical context.

This paper will not advocate a single approach or methodology for applying and evaluating search and metadata to a digital representation; rather, it will argue that digital historians should *think archivally* when considering how these components contribute to a representation's historical contextualization. Refinement of this mindset through rigorous,

systematic, and interdisciplinary theoretical and practical experimentation could benefit scholarship, peer review, pedagogy, public history, and cultural heritage.⁶⁷

References

- Archives and Public History Digital.* <http://aphdigital.org/> (accessed 12 March 2010).
- Arthur, P.** (2008). 'Exhibiting History: The Digital Future'. *reCollections: The National Museum of Australia*. **3(1)**. http://recollections.nma.gov.au/issues/vol_3_no_1/papers/exhibiting_history/.
- Ayers, E. L.** (2002). 'Technological Revolutions I Have Known'. *Computing in the Social Sciences and Humanities*. Burton, O.V. (ed.). University of Illinois Press, pp. 19-28.
- Cook, T.** (2001). 'Archival Science and Postmodernism: New Formulations for Old Concepts'. *Archival Science*. **1**: 3-24.
- Duff, W. and Harris, V.** (2002). 'Stories and Names: Archival Description as Narrating Records and Constructing Meanings'. *Archival Science*. **2**: 263-285.
-
- Notes**
1. Terry Cook. "Archival science and postmodernism: new formulations for old concepts." *Archival Science*. 1, 2001. 22.
 2. The term "digital historiography" has lingered in the background of the field throughout the last decade, most notably in a series of reviews by David Staley in the *Journal of the Association for History and Computing* between 2001-2003.
 3. For a recent survey of the digital history field and its variant representational forms see Paul Arthur. "Exhibiting history: The digital future." *reCollections: The National Museum of Australia*. Vol. 3, number 1. http://recollections.nma.gov.au/issues/vol_3_no_1/papers/exhibiting_history/. Accessed March 12, 2010.
 4. Edward L. Ayers. "Technological Revolutions I Have Known." In *Computing in the Social Sciences and Humanities*. Ed. Orville Vernon Burton. University of Illinois Press, 2002. 27. My call for a rigorous digital historiography coincides with Ayers' own remarks, when he writes: "Whatever a project's scale and level of complexity, new media should meet several standards to justify the extra effort they take to create, disseminate, and use."
 5. For further discussion on how archival description can shape the narrative embedded within archival records, see Wendy Duff and Verne Harris. "Stories and Names: Archival Description as Narrating Records and Constructing Meanings." *Archival Science*. 2: 263-285, 2002.
 6. Among the possible applications could be the development of higher education curriculum constructed around a hybrid

digital history-archival studies model. NYU's Archives and Public History is one of the leading programs that have taken up the call to teach archival theory alongside digital history theory and practice. It recently unveiled a new website showcasing its revamped academic program: <http://aphdi.gital.org/>. Accessed March 12, 2010.

7. The thoughts and ideas expressed in this abstract and the conference presentation are entirely my own and do not necessarily reflect those of NEH or any other federal agency.

e-Vocative Cases: Digitality and Direct Address

Lisa Swanstrom

swanstro@gmail.com

HUMlab, Umeå University, Sweden

Electronic literature poses several exciting challenges and questions to literary scholars: How do we balance our interpretations of digitally born works against the specific modes of production that make such works possible? How do our conceptions of authorial intent shift in relation to works that solicit active participation from their readers? How do we account for readers' participation in such works, as well as the way their experiences shape and re-shape the text? In this paper I offer one strategy of interpretation that cuts across some of these questions: tracing the path of direct address in works that are digitally born, a technique that both emerges and departs from conventional literary practice.

When I visit a certain website, I am greeted in a peculiar fashion: an animated avatar with a human form speaks to me, blinks at me, and follows my mouse movements on the screen with her eyes while I read. On another site, a string of text hails me and addresses me by name, purporting to welcome me to all the treasures contained within its digital domains. As startling as these salutations initially seemed and as commonplace as they have become, I remain intrigued by their overt and shameless invocation of the reader — in this case, me.

Strictly speaking, this mode of address should not be possible, at least not according to the familiar conventions of literary tradition. In *Anatomy of Criticism*, Northrop Frye states the matter unequivocally: "Criticism can talk, and all the arts are dumb...there is a most important sense in which poems are as silent as statues. Poetry is a disinterested use of words: it does not address a reader directly" (4). While the examples of address above are decidedly not the poetical specimens Frye has in mind, his stance nevertheless serves as a firm response to a larger problem, one that has endured since the time of Socrates and persists to this day, a problem that can be crudely summarized in the following terms: there has always been something of a gap between the written word and its reception.

Each time I see my own name staring back at me, however, I question whether the gap between text and reader has been in some way bridged, or at least contracted. Each time the avatar speaks to me, I am unable to locate myself in relation to the text no

matter which paradigm I might use to explicate our relationship. Within a spectrum bounded at one end by the New Critical emphasis on textual autonomy and at the other by the “virtual” text that emerges necessarily as a correspondence between author and audience in reader-response theory, I do not know where I stand. With the announcement of my own name, I am aware that I have been identified, and therefore can no longer even maintain the convenient illusion of being, as a reader, either ideal or implied. I have been specified. The “text,” such as it is, has called me out.

The spectrum I have identified here is, of course, absurdly streamlined and unequally weighted. The New Critics exclude the reader’s thoughts as a given principle, while reader-response theory alone has perhaps generated more ways of labeling its reading audience than the sum of other critical interventions combined — in addition to offering a strong and convincing counterpoint to Cleanth Brooks’ ideal reader, Wolfgang Iser’s implied reader is only one star in a constellation of terms that includes the mock reader, the actual reader, the fictionalized reader, the hypothetical reader, the narrative reader, the ideal narrative reader, and the “real” reader, not to mention Stanley Fish’s interpretive reading communities (Brooks, 24; Iser, *The Implied Reader*, xii; Rabinowitz, 125-128; Fish, 219).

In all of these models of reception, the impulse to name the reader, to re-assert her importance in the construction of textual meaning, still participates in the tacit agreement that this reader, whoever she may be, is never fully concretized by the written text. How could she be? Rather, a “virtual” text emerges as a sort of ghostly correspondence between the two, one that is nigh on impossible to trace. In the words of Iser, “It’s difficult to describe this interaction...because...of course, the two partners in the communication process, namely, the text and the reader, are far easier to analyze than is the event that takes place between them” (“Interaction Between Text and Reader,” 107).

The hypothesis that I would like to test in this essay is that works of electronic literature push the issue of responsibility and specificity into uncharted readerly terrain. What if, in certain examples of electronic literature, direct address online were specific to you, the reading reader, and not an implied reader? Put more specifically, pressed even further, what if direct address online were to make traceable the ghostly correspondence between reader and text that Iser outlines? This is not as far-fetched as it seems. As we shall see, the reader’s participation in some examples of electronic literature is required for textual constitution in ways that are fundamentally different from even the most successful and extreme examples of non-linear narrative practices found in print. In the case of electronic literature, direct

address functions to bring the text into being, by signaling the reader and requiring a *response* of her. Even more remarkable, this response has the ability to become a part of the initial text, such that the text that emerges is literally constituted through the feedback that exists between the reader’s actions and the author’s words.

While many claims about interactivity and customization have been made about electronic literature, there has not yet been a sustained attempt to consider the more specific mode of address that occurs in such works in relation to overtly literary practice. In the space that follows I attempt to remedy this by considering instances both subtle and overt that occur in select works of electronic literature — including Dan Waber and Jason Pimble’s “I, You, We,” Mary Flanagan’s [theHouse], and Emily Short’s “Galatea” — that signal, cue, or otherwise point outside themselves to the reader as she progresses through the text. If the use of the vocative in conventional literary texts has the ability to point not only to characters within their narrative confines, but to an entire social, political and cultural discourse that lies tantalizingly close, yet perhaps ultimately outside the textual boundaries, I explore whether modes of address in online works allow us to exceed these boundaries altogether.

References

- Brooks, Cleanth** (2004). 'The Formalist Critics'. *Literary Theory: an Anthology*. Julie Rivkin, Michael Ryan (eds.). Malden, MA: Blackwell.
- Fish, Stanley** (2004). 'Interpretive Communities'. *Literary Theory: an Anthology*. Julie Rivkin, Michael Ryan (eds.). Malden, MA: Blackwell.
- Frye, Northrop** (1957). *Anatomy of Criticism*. Princeton: Princeton UP.
- Iser, Wolfgang** (1980). 'Interaction Between Text and Reader'. *The Reader in the Text: Essays on Audience and Interpretation*. Suleiman, Crosman (eds.). Princeton: Princeton UP.
- Rabinowitz, Peter J.** (Autumn, 1977). 'Truth in Fiction: A Reexamination of Audiences'. *Critical Inquiry*. 4:1: 121-141.

Digitizing the Act of Papyrological Interpretation: Negotiating Spurious Exactitude and Genuine Uncertainty

Tarte, Ségolène M.

segolene.tarte@oerc.ox.ac.uk
University of Oxford, UK

1. Digitization is a sampling process

The act of papyrological interpretation is a continuous thought process that unravels non-linearly (Youtie, 1963; Terras, 2006). Throughout this sense-making process, ancient and scarcely legible documents progress from the status of pure physical objects to that of meaningful historical artefacts. Within the e-Science and Ancient Documents project,¹ we aim to make explicit some of the implicit mechanisms that contribute to the development of hypotheses of interpretation by designing and implementing a web-based software offering digital support for the hermeneutic task. This tool aims to record the intermediary hypotheses of interpretation, thus keeping track of the rationale and allowing easier and better revision when required. The model we have adopted (Roued Olsen et al., 2009) is that of a network of percepts, where a percept is defined as a minor interpretation that stems from perception and cognition (Tarte, 2010). An understanding of expert knowledge and of how it is mobilised is required to identify the crucial steps that allow us to reconstruct a rationale. The level of the granularity at which we choose to provide support also is essential to the usability of the software. Further, each percept, each intermediary interpretation, each piece of evidence used either to support or to invalidate a claim is potentially mutable. The implementation of an Interpretation Support System (ISS) taking these considerations into account poses the question of how to digitize or record a thought process; it is an epitome of the ‘continuous-to-discrete’ (or ‘analogue-to-digital’) problem. In the theoretical and life sciences, measurement devices are developed to sample the signals of interest. Then, based on the discrete sampled signal, on an underlying model of the behaviour of the signal, and on more general knowledge of signal processing and information theory (e.g. the Nyquist-Shannon

sampling theorem), the continuous signal can be reconstructed with minimal deviation from the original signal. Similarly, the ambition of our ISS is, based on an appropriate model of the papyrological hermeneutic task, to allow the user to capture the information necessary to the reconstruction of the rationale that yielded a given interpretation. Two difficulties in sampling the interpretive thought process are: (1) to take advantage and to beware of the sense of scientific rigour that digitization conveys; and (2) to allow the digital expression of uncertainty and mutability of percepts.

In this paper, we explain how, while attempting to digitize the papyrological interpretation act, we strive to avoid spurious exactitude and accommodate genuine uncertainty.

2. Choosing what to digitize and how

The papyrological model of reading developed by Terras (Terras, 2006) identified ten levels of reading, corresponding to ten levels of granularity at which an interpretation in progress is discussed. Tools stemming from web-based technology (Bowman et al., 2010), image processing (Tarte et al., 2009) and artificial intelligence can help support the digitization of the hermeneutic task (see fig. 1). To illustrate how we negotiate between spurious exactitude and genuine uncertainty, we focus here on two specific stages of the digitization of the papyrological interpretation process: how artefact digitization is being performed; and how, by identifying the mechanisms that trigger the jumps between the ten levels of reading, we propose to address the representation of uncertainty.

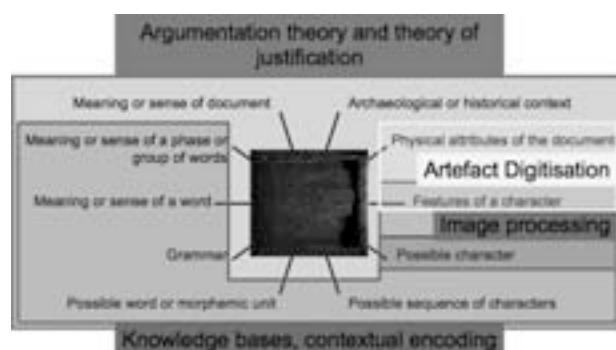


Figure 1: Model of the act of interpretation detailing the various levels of reading based on (Terras, 2006) and the tools involved in the implementation of our ISS.

2.1. Digitizing the text-bearing artefact

The problem of spurious exactitude is most prevalent at the stage where the text-bearing artefact is digitized. For stylus tablets, for example, high-

resolution pictures are not enough. The materiality of the artefact needs to be taken into account in a similar way as the experts exploit it in the real-world. The guiding principle we choose to follow in this context is mimesis. Indeed, when the papyrologists have physical access to such an incised tablet, in order to see better the incised text, they lay the tablet flat on their hand, lift it at eye level and expose it to raking light while applying pitch-and-yaw motions to it. This signal enhancement strategy exploits the shadow-stereo principle by which stronger and more mobile shadows and highlights occur at the incisions than they do on the bare wood; the text is thereby accentuated. Digitally imitating this process, we capture series of images of incised tablets with varying light positions (Brady et al., 2005), allowing users to reproduce digitally the visual phenomenon they naturally exploit in the real-world. Note that, similarly to signal measurement devices, we adopt a digitization process that is already part of the interpretation process. And the intention behind artefact digitization, as well as the intention behind signal measurement, is always an implicitly set variable that affects downstream results.

2.2. Digitizing the thought process

When attempting to capture the milestones of the thought process that builds an interpretation of an ancient or damaged text, we need to capture the uncertainty of the intermediary percepts and their mutability. A numerical approach to uncertainty such as bayesian networks could have been adopted, but such a quantification of uncertainty usually presupposes that problems are complete, i.e. that all the alternatives to a given situation are known (Parsons & Hunter, 1998). Instead, we have decided to turn to argumentation theory (Parsons & Hunter, 1998) and theory of justification (Haack, 1993), and combine them to provide a formal, yet invisible, epistemological framework that will allow us to point out inconsistencies without forbidding them. Indeed, inconsistencies in an unravelling interpretation naturally occur and can be rooted either in the implicit expectations of the user or in the validity of the actual claims (see Tarte, 2010 for an account of an inconsistency due to an implicit assumption and its resolution). The balance to be found here (Shipman III & Marshall, 1999) is between on the one hand the usefulness of a formal system as a backbone to support reasoning under uncertainty and make implicit mechanisms explicit, and on the other the excess of formalism and explicit formulation that can become a hindrance by creating for the user an overhead disruptive to the interpretation process. Here again, our design choice is based on the observation of the experts at work. We allow both palaeographical and philological approaches in combination, through the possibility of tracing the

letters and of handling the text as a crossword puzzle (Roued-Cunliffe, 2010); these are both approaches that we have identified as the main strategies experts develop when interpreting documents (Tarte, 2010). The expression of uncertainty is then made inherent to the mode of interaction with the artefact, and the transposition of the real-world tasks of drawing and crossword puzzle solving allows us to keep the interface intuitive while, in the background, more formal mechanisms can run routines such as consistency checks and consultation and constitution of knowledge bases. In her doctoral work, Roued-Cunliffe (Roued-Cunliffe, 2010) is currently concentrating on the crossword puzzle approach and combining it to consultation of knowledge bases through web-services.

3. Conclusion: digitization is also interpretation

Digital technologies can easily trick the mind into thinking that their use confers an exactitude on the results obtained with their support. It is however worth noting that in the sciences too, digitization is always made with an intention. When looking to sample a continuous signal, be it a temperature as a function of time, or a thought process as a function of time, the sampling strategy is always adopted in the light of an intention. Digitization is actually also an act of interpretation. To record digitally the continuous papyrological interpretation process, we have to identify clearly our final aim, and to adapt our sampling strategy accordingly. Here, our aim is to enable to record, reconstruct, back-track if necessary, the interpretation process by making explicit (some of) the epistemological evidence substantiating the interpretation in progress; an added benefit to the software is that it will also enable easier production of an edition of a text, as the evidence will have been laid out clearly. Capturing uncertainty is vital to the recording process, and being conscious that its very capture is also part of the hermeneutic task is crucial to allow the software design to take on board the elements that are core to the whole interpretation process.

Acknowledgements

This work was supported by the joint AHRC-EPSRC-JISC Arts and Humanities e-Science Initiative UK [grant AH/E00654X/1].

The author wishes to thank Prof. Alan Bowman, Prof. Sir Michael Brady, Dr. Roger Tomlin, Dr. Melissa Terras, Dr. Charles Crowther and Henriette Roued-Cunliffe for their support, as well as the reviewers for their helpful comments.

References

- Bowman, A. K., Crowther, C. V., Kirkham, R., Pybus, J.** (2010). 'A virtual research environment for the study of documents and manuscripts'. *Digital Research in the Study of Classical Antiquity*. Bodard, G., Mahony, S. (eds.). London: Ashgate Press.
- Brady, M., Pan, X., Schenck, V., Terras, M., Robertson, P., Molton, N.** (2005). 'Shadow stereo, image filtering and constraint propagation'. *Images and Artefacts of the Ancient World*. Bowman, A. K., Brady, M. (eds.). British Academy Occasional Paper. Oxford: Oxford University Press/British Academy. V. 4, pp. 15–30.
- Haack, S.** (1993). 'Double-aspect foundherentism: A new theory of empirical justification'. *Philosophy and Phenomenological Research*. 53(1): 113–128.
- Parsons, S., Hunter, A.** (1998). 'A review of uncertainty handling formalisms'. *Applications of Uncertainty Formalisms*. Carbonell, J. G., Siekmann, J. (eds.). Lecture Notes in Artificial Intelligence (Lecture Notes in Computer Science). Berlin / Heidelberg: Springer. V. 1455, pp. 8–37.
- Roued-Cunliffe, H.** (2010: forthcoming). 'Towards a Decision Support System for Reading Ancient Documents'. *Literary and Linguistic Computing*.
- Roued Olsen, H., Tarte, S. M., Terras, M., Brady, J. M., Bowman, A. K.** (2009). 'Towards an interpretation support system for reading ancient documents'. *Digital Humanities'09*. 2009, pp. 237–39.
- Shipman III, F. M., Marshall, C. C.** (1999). 'Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems'. *Computer Supported Cooperative Work*. 8: 333–52.
- Tarte, S. M.** (2010: forthcoming). 'Papyrological investigations: Transferring perception and interpretation into the digital world'. *Literary and Linguistic Computing*.
- Tarte, S. M., Wallom, D., Hu, P., Tang, K., Ma, T.** (2009). 'An image processing portal and web-service for the study of ancient documents'. *2009 Fifth IEEE International Conference on e-Science*. 2009.
- Terras, M.** (2006). *Image to Interpretation. An Intelligent System to Aid Historians in Reading the Vindolanda Texts*. Oxford: Oxford University Press.
- Youtie, H. C.** (1963). 'The papyrologist: artificer of fact'. *Greek, Roman and Byzantine Studies*. 4(1): 19–33.

Notes

1. Project website: <http://esad.classics.ox.ac.uk/>

Reimagining the Dictionary, or Why Lexicography Needs Digital Humanities

Tasovac, Toma

ttasovac@humanistika.org

Center for Digital Humanities (Belgrade), Serbia

The promise of eLexicography stems not only from the transformation of the production medium, but also from the technological feasibility of representing linguistic complexity. Even though modern lexicography is unimaginable without computer technology (Hockey, 2000a; Knowles, 1989; Meijis, 1992), the sheer use of computers in producing a dictionary or delivering it electronically does not automatically transform a dictionary from "a simple artefact" to a "more complex lexical architecture," to use Sinclair's (2000) formulations.

Calling dictionaries "simple artefacts" is itself a rhetorical oversimplification: there is certainly nothing simple about a dictionary — whether we look at it as a material object, cultural product or a model of language. Yet the overall structure of dictionaries as extended word lists has not changed in centuries (Hausmann et al., 1989; Fontenelle, 2008; Atkins and Rundell, 2008). Admittedly, a great deal of factual information is packed into a prototypical lexicographic entry, but a defined term often remains in isolation and insufficiently connected or embedded into the language system as a whole. This is what Miller refers to as the "woeful incompleteness" (Miller et al.) of a traditional dictionary entry, and what Shvedova sees as its "paradoxical nature" — dictionary entries tend to be "lexicocentric" while language itself is "class-centric" (Шведова, 1988).

Furthermore, the advances in digital humanities, textual studies and postmodern literary theory do not seem to have had a profound effect on the way we theorize or produce dictionaries. Surely, many important lexicographic projects have been digitalized and gone online; web-portals increasingly offer cumulative searches across different dictionaries; and eLexicography is a thriving field (Lemberg et al., 2001; Hockey, 2000a; de Schryver; Hass, 2005; Nielsen, 2009; Rundell, 2009; Hass, 2005), yet dictionaries — often commercial enterprises which are guided by predominantly economic concerns — remain by far and large discrete objects: no more and no less than digitalized versions of stable, print editions. We still

consult dictionaries by going to a particular web site. Dictionaries do not come to us.

The time is ripe to ask — both in theoretical and practical terms — a new set of questions: how has the electronic text changed our notion of what a dictionary is (and ought to be); how have the methods of digital humanities and the advances made in digital libraries altered our idea of what a dictionary can (and should) do? And, finally, where do we go from here?

The dictionary is a kind of text. In print culture, the dictionary, like every other text, had its material and semantic dimension. The semantic dimension was represented on its visible surface, whereas its depth was in the mind of the reader, or what Eco refers to as the "encyclopedia of the reader." (Eco et al., 1992; Eco, 1979). Yet if we — as we should — start thinking of the dictionary as a kind of electronic text, the way Kathrine Hayles and others have done for electronic literature, we will have no choice but to strip the dictionary of its finality and its "object-ness" and see in it, instead, only one possible manifestation of the database in which it is stored (Hayles, 2003; Hayles, 2006; Folsom, 2007). A digital text can be not only edited, transformed, cut and pasted — as part of our computational textual kinetics — but is always part of other activities: search, downloading, surfing. In other words, an electronic text is unimaginable without its context (Aarseth, 1997; DeRose et al., 1990; Hockey, 2000b).

The dictionary, then, should be seen as a kind of semantic potential that can be realized through its use. But in order to truly fulfill this potential, the dictionary needs to be embedded in the digital flow of our textual production and reception. That is why we cannot think of dictionaries any more without thinking about digital libraries and the status which electronic texts have in them (Andrews and Law, 2004; Candela et al., 2007; Kruk and McDaniel, 2009; Maness, 2006; Miller, 2005; Novotny, 2006). To be truly useful for any kind of textual studies, the digital library must "explode" the text (by providing full-content searchability, concordances and indexes, metadata, hyperlinks, critical markup etc.) instead of "freezing" it as an image, which, albeit digital, is computationally neither intelligible nor modifiable as text. In smart digital libraries, a text should not only be an object but a service; not a static entity but an interactive method (Tasovac, forthcoming). The text should be computationally exploitable so that it can be sampled and used, not simply reproduced in its entirety. This kind of atomic approach to textuality poses a host of challenges (legal, ethical, technical and intellectual, to name just a few), but it opens up the possibility of creative engagement with the digital text in literary studies (text mining, statistical text

comparison, data visualization, hypertextual systems etc.).

The consequence of this "explosive" nature of the electronic text is of paramount importance for eLexicography and the reformulation of the dictionary not as an object, but a service. We should start thinking of and building dictionaries as fully embeddable modules in digital libraries, or, to put it differently, build digital libraries which integrate dictionaries as part of their fundamental infrastructure and allow an ever-expandable process of associating words in an electronic text with an equally changeable record in a textual database. The changeability of the dictionary entry will, in turn, defer *ad infinitum* the notion of a particular dictionary edition — other than as temporary snapshot of the database. The dictionary as an evolving process will be in a permanent beta state.

The future of electronic dictionaries undoubtedly lies in their detachability from physical media (CD, DVD, desktop applications) and static locations (web portals). If we think of the dictionary as a service with an API¹ that can be called from any Web page, we can actually start thinking about any (electronic) text as a direct entry point to the dictionary. If every word in a digital library is a link to a particular entry in the dictionary, electronic textuality as such becomes an extension of lexicography: the text begins to contain the dictionary in the same way that the dictionary contains the text.

The Center for Digital Humanities (Belgrade, Serbia) is putting these theoretical considerations into practice while working on its flagship *Transpoetika Project* (Tasovac, 2009). *Transpoetika* (see Figure 1) is a collaborative, class-centric, bilingualized Serbian-English learner's dictionary based on the architecturally complex, machine-readable semantic network of the Princeton Wordnet (Fellbaum, 1998; Vossen, 1998; Stamou et al., 2002; Tufis et al., 2004). It is part of a scalable, web-based, digital framework for editing and publishing annotated, fully-glossed study editions of literary works in the Serbian language, primarily aimed and students of Serbian as a second or inherited language.

Transpoetika has been designed to be deployed as a web service and therefore linked from and applied to a variety of textual sources online. Portions of the project, such as the Serbian Morpho-Syntactic Database (SMS) already function as a web service internally and will also be made public and free once the sufficient funding for the project has been secured. *Transpoetika* can also interact with other web services: by using Flickr as a source of illustrations, and Twitter as a source of "live quotes" in the entries, the *Transpoetika* Dictionary explores the role of serendipity in a lexicographic text.

The overarching goal of the Belgrade Center for Digital Humanities (CDHN) is to produce a pluggable, service-based, meta-lexicographic platform for the Serbian language, which will interact with various Web-based digital libraries, and contain not only our own bilingualized Serbian Wordnet, but also historical Serbian dictionaries that the CDHN is digitalizing, such as, for instance, the classic Serbian-German-Latin Dictionary by Vuk Stefanović-Karadžić (1818 and 1852). The platform could, in theory, be extended to include and consolidate a number of other, more specialized, lexicons. This is, in any case, the general direction we would like to take.

I would like to conclude with a *hysteron-proteron*, which, in Samuel Johnson's Dictionary of the English language was defined as "a rhetorical figure: when that is last said, which was first done." From the very beginning of this paper, I spoke of the dictionary, which every careful reader would have marked as a serious lexicographic faux-pax. There is and never was such a thing as a singular and uniquely authoritative source of information about words and their meanings. There is no such thing as the (Platonic, ideal) dictionary but rather a myriad manifestations of its imagined hypertextual prototype. I believe, nonetheless, that we should, in the digital age and with the ongoing developments of the digital humanities, reclaim the dusty notion of the dictionary and boldly, though not without self-irony, keep trying to imagine what that "thing" — the dictionary — could be. If only with the goal of making it — in its traditional, leather-bound, sense — completely obsolete.



Figure 1

References

- Aarseth, E.J.** (1997). *Cybertext: Perspectives on Ergodic Literature*. Baltimore, MD: The Johns Hopkins University Press.
- Andrews, J., Law, D.G. (eds.)** (2004). *Digital Libraries: Policy, Planning, and Practice*. Aldershot, Hants, England; Burlington, VT: Ashgate.

- Atkins, B.T.S., Rundell, M.** (2008). *The Oxford Guide to Practical Lexicography*. Oxford; New York: Oxford University Press.
- Candela, L., et al.** (2007). 'The DELOS Digital Library Reference Model: Foundations for Digital Libraries'. Version 0.98. http://www.delos.info/index.php?option=com_content&task=view&id=345.
- de Schryver, G.-M.** (2003). 'Lexicographer's Dreams in the Electronic-Dictionary Age'. *International Journal of Lexicography*. 143-199.
- DeRose, S., et al.** (1990). 'What Is Text, Really?'. *Journal of Computing in Higher Education*. 1: 3-26.
- Eco, U.** (1979). *The Role of the Reader: Explorations in the Semiotics of Texts*. Bloomington: Indiana University Press.
- Eco, U., et al.** (1992). *Interpretation and Overinterpretation*. Cambridge; New York: Cambridge University Press.
- Fellbaum, C.** (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.
- Folsom, E.** (2007). 'Database as Genre: The Epic Transformation of Archives'. *PMLA*. 122: 1571-1579.
- Fontenelle, T.** (2008). *Practical Lexicography: A Reader*. Oxford; New York: Oxford University Press.
- Hass, U. (ed.)** (2005). *Grundfragen der elektronischen Lexikographie : Elexiko, das Online-Informationssystem zum deutschen Wortschatz*. Berlin; New York: W. de Gruyter.
- Hausmann, F.J., Reichmann, O., Wiegand, H.E. (eds.)** (1989). *Wörterbücher: ein internationales Handbuch zur Lexikographie*. Berlin; New York: W. de Gruyter.
- Hayles, N.K.** (2003). 'Deeper into the Machine: The Future of Electronic Literature'. *Culture Machine*. 5.
- Hayles, N.K.** (2006). 'Traumas of Code'. *Critical Inquiry*. 33: 136-157.
- Hockey, S.M.** (2000a). 'Dictionaries and Lexical Databases'. *Electronic texts in the humanities: Principles and Practice*. Oxford; New York: Oxford University Press, pp. 146-171.
- Hockey, S.M.** (2000b). *Electronic texts in the humanities: Principles and Practice*. Oxford; New York: Oxford University Press.
- Knowles, F.E.** (1989). 'Computers and Dictionaries'. *Wörterbücher: ein internationales Handbuch zur Lexikographie*. Hausmann, F.J., Reichmann, O., Wiegand, H.E. (eds.). Berlin; New York: W. de Gruyter, pp. 1645-1672.
- Ryszard Kruk S., McDaniel B. (eds.)** (2009). *Semantic Digital Libraries*. Berlin: Springer.
- Lemberg, I., Schröder, B., Storrer, A. (eds.)** (2001). *Chancen und Perspektiven computergestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher*. Tübingen: M. Niemeyer.
- Maness, J.M.** (2006). 'Library 2.0 Theory: Web 2.0 and Its Implications for Libraries'. *Webology*. 2.
- Meijs, W.** (1992). 'Computers and Dictionaries'. *Computers and Written Texts*. Butler, C. (ed.). Oxford; Cambridge, Mass.: Blackwell, pp. 141-165.
- Miller, G.A. et al.** 'Introduction to WordNet: An Online Lexical Database'. *Five papers on WordNet*.
- Miller, P.** (2005). 'Web 2.0: Building the New Library'. *Ariadne*. 45.
- Nielsen, S.** (2009). 'Reviewing printed and electronic dictionaries: A theoretical and practical framework'. *Lexicography in the 21st Century. In honour of Henning Bergenholz*. Nielsen, S., Tarp, S. (eds.). Amsterdam: John Benjamins, pp. 23-41.
- Novotny, E.** (2006). *Assessing Reference and User Services In a Digital Age*. Binghamton, NY: Haworth Information Press.
- Rundell, M.** (2009). 'The future has arrived: a new era in electronic dictionaries'. *MED Magazine*. 54.
- Sinclair, J.** (2000). 'Lexical Grammar'. *Darbai ir Dienos*. 24.
- Stamou, S., et al.** (2002). 'BALKANET A Multilingual Semantic Network for the Balkan Languages'. *Proceedings of the International Wordnet Conference*. Mysore, India, 21-25 January 2002, pp. 21-25.
- Tasovac, T.** (2008). 'Why not every picture is worth a thousand words: digital libraries from a textual perspective'. *Proceedings of the International Conference "Electronic Libraries - Belgrade 2008"*. University of Belgrade, 25-27 September 2008.
- Tasovac, T.** (2009). 'More or Less than a Dictionary: WordNet as a Model for Serbian L2 Lexicography'. *Infoteka - Journal of Informatics and Librarianship*. 10: 13a - 22a.
- Tufis, D., Cristea, D., Stamou, S.** (2004). 'BalkaNet: Aims, Methods, Results and Perspectives. A General Overview'. *Science and Technology*. 7: 9-43.
- Vossen, P. (ed.)** (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht, The Netherlands; Boston, Mass.: Kluwer Academic.
- Шведова, Н.** (1988). 'Парадоксы словарной статьи'. *Национальная специфика языка и ее отражение в нормативном словаре. Сборник*

статьей. Караполов, Ю.Н. (ed.). Москва: Наука, pp. 6-11.

Notes

1. The first publicly available dictionary application programming interface was made available by the Wordnik project in October 2009. See <http://api.wordnik.com/signup/>.

Contexts, Narratives, and Interactive Visual Analysis of Names in the Japanese Hyohanki Diary

Toledo, Alejandro

alex@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

Visualization tools for the discipline of history are being used increasingly. Historians can explore their efficacy as both educational instruments and platforms for displaying research findings. In order to obtain patterns of significance, historian's work involves manipulating, memorizing, and analyzing substantial quantities of information from the series of documents at their disposal. In so doing, a central purpose emerges: the construction of a narrative that best fits into the representation of the past.

Visualization tools have proved to be effective for facilitating users' analytical tasks (Heer et al., 2008; Heer et al., 2009; Weizhong and Chen, 2007, James and Cook, 2005). Moreover, they have proved to be fruitful in the context of the digital humanities (Bonnet, 2004; Dalen-Oskam and Zundert, 2004). Recently, we proposed a visualization system for analyzing aristocrat names in a Japanese historical diary called Hyohanki (Toledo et al., 2009). In our system, the stacked graph is utilized to analyze the time series of those names. Stacked graphs, stacking time series on top of each other, are a useful method to time series visualization, resulting in a visual summation of time series values that provides an aggregate view stratified by individual series. Projects such as NameVoyager (Wattenberg et al., 2005) and sense.us (Heer et al., 2009) used animated stacked graphs to explore demographic data.

In this paper, to preserve contexts through the stacked graph usage, we propose an extension of our previous work. Our system provides two functionalities: an interaction control for saving, querying, and deleting views, as well as a dynamic repository of views representing the context of the stacked graph usage. Keeping useful contexts facilitates coherent narratives (Bonnet, 2004). To

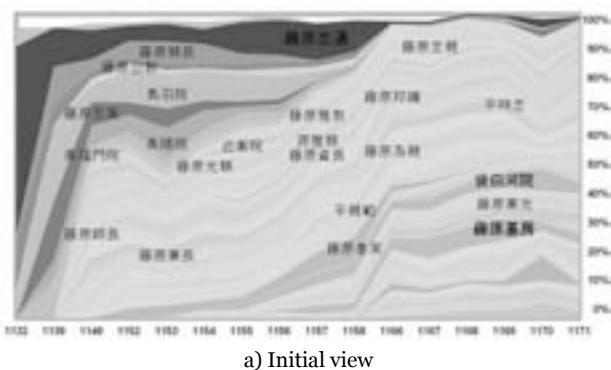
support historians' efforts in keeping contexts through interactive visual analysis of time series, both functionalities aim at recording how users receive units of information appropriated to construct narratives.

To analyze the end-user perception of our system, we conducted a heuristic evaluation (Nielsen, 1992) with a domain expert who explored the data using the tool. This heuristic evaluation was a form of user study in which the expert reviewed the system to suggest advantages and disadvantages against the new functionalities. This approach helps to further elucidate the requirements and how the system meets experts' needs. The results provide useful guidance for highlighting known historical facts as well as hints to unknown historical facts.

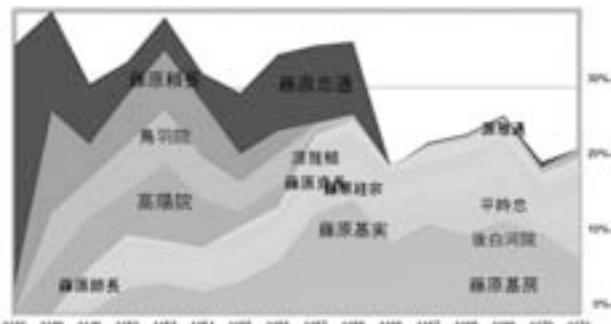
1. Methodology

Our visualization system is based on a time series set containing the quantitative analysis of name occurrences in the Hyohanki diary. The set contains 121 time series in a timescale spanning from 1132 to 1171. Some parts of the diary suffer from missing data; for that reason, that period includes only the years 1132, 1139, 1149, 1152-1158 and 1166-1171. Likewise, the data has been normalized in order to measure the percentage relative value of the number of occurrences of a given name, in a given year, with the total percentage of that year. Additionally, using the Euclidean distance metric, we calculated the trends' similarity between time series. For each name, we recorded their five most similar trends.

The method used to visualize the data is straightforward. Given a set of aristocrat names' time series, a stacked graph is produced (Fig. 1a). The x axis corresponds to year and the y axis to the occurrence ratio — in percentage, as the data has been normalized —, for all names currently in view. The stacked graph contains a set of stripes, each one representing a name. The width of each stripe is proportional to the ratio of that name mentioned in a given year. The stripes are colored blue and the brightness of each stripe varies according to the number of occurrences, so that the most mentioned names during the whole period are darkest and stand out the most. Likewise, the name's font size follows a similar encoding; the higher the occurrence frequency of a name, the larger its font size.



a) Initial view



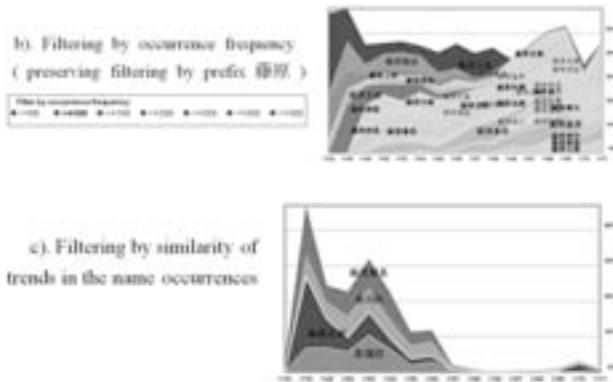
b) Filtered view with occurrence frequency above 200

Fig. 1 Screenshots of the stacked graph¹

The interaction capability provided by our visualization system comes in two flavors. First, a *stripe filtering* mechanism allows users to filter the stacked graph in three ways: prefix typing, occurrence frequency, and stripe similarity. The second interaction control, which we call contextualization keeping, aims at keeping the views produced by users as they interact with the visualization system. Because stripe filtering acts as a view producer, and *contextualization keeping* as a dynamic repository of those views, we believe that both types of interaction controls embody a useful tool for visual exploratory analysis.

Stripe filtering forms the base for the source of views. Consider for example the interaction control for filtering by string prefixes (Fig. 2a). Users are allowed to type either a complete name or only a prefix of that name. The figure shows a view corresponding to the prefix 藤原 (family name Fujiwara). The number of possible views that this interaction control can produce is equivalent to at least the size of the time series set. The Hyohanki diary produces about 150 views from a total of 121 names.





In addition to the stripe filtering, our system provides filtering by occurrence frequency (Fig. 2b), which uses a set of discrete values leading to seven possible views. Users, however, will likely want to combine the effects of more than one interaction control; for which case, the amount of possible views might increase enormously. An extreme scenario may occur when considering the last type of filtering, similarity of trends (Fig. 2c), which allows users to reveal those stripes that are similar to others. To cope with this issue, we propose in this paper a preliminary implementation for keeping the contextualization of the stacked graph usage.

Our proposal for keeping contextualization in the stacked graph is composed of two parts: a toolbar which users can use to save and delete views and a repository of the resulting views from the effects of the toolbar usage (Fig. 3). For each saved view, the system records contextual information of the current visualization state. This information is saved in terms of the stripe filtering interaction controls used to produce those views. In order to support their sense making tasks, users can operate on the toolbar, especially on the saving button, whose effect suggests a mapping between the intrinsic facts of the view with the user mental model created from that view. The repository, on the other hand, provides the stacked graph with the ability to serve as a cognitive tool. This is because users can recover views from the repository, reason on them depending on their own criteria, and decide whether to keep the view or delete it. The main attraction of our proposal is its potential provision of stacked graph's usage paths as the results of a visual exploratory analysis.

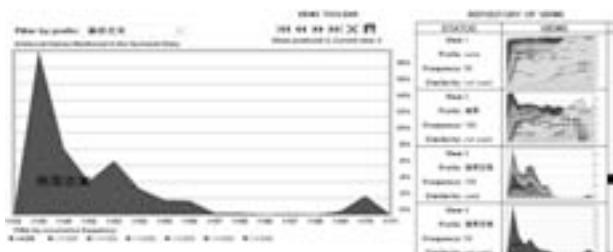


Fig. 3 Interaction control for keeping contextualization in the stacked graph

2. Results and Discussions

This work attempts to provide a stacked graph module for gathering views to construct narratives. A concrete prototype has been developed and evaluated with an expert. In general, our user had a favorable experience with the system, though he commented that the similarity filtering was confusing and should provide a more effective approach for this. Below we have an observation the expert was able to make using the tool.

Fig. 3 shows a use case of our visualization system. Six views were saved into the repository using a deductive reasoning approach, i.e., going from the stacked graph as a whole to a view containing only one name. From those views, only three of them gained relevance to construct the narrative, i.e., views 2-4. As shown in Fig. 4, or View 2 of Fig. 3, there are no significant variations in the trends with prefix 藤原 (family Fujiwara) — to some extent they were mentioned over the same period —. However, the similarity view of 藤原忠実 (Fujiwara no Tadazane) in View 3 of Fig. 3 (see also Fig. 2c) excludes his eldest son 藤原忠通 (Fujiwara no Tadamichi). In contrast, Fujiwarano Yorinaga (藤原頼長) not only is a member of that group, but he has the most similar trend to his father (Tadazane). This situation was used to construct the following narrative: "*Tadazane was possibly closer to his second son*", which actually confirms a known historical fact.

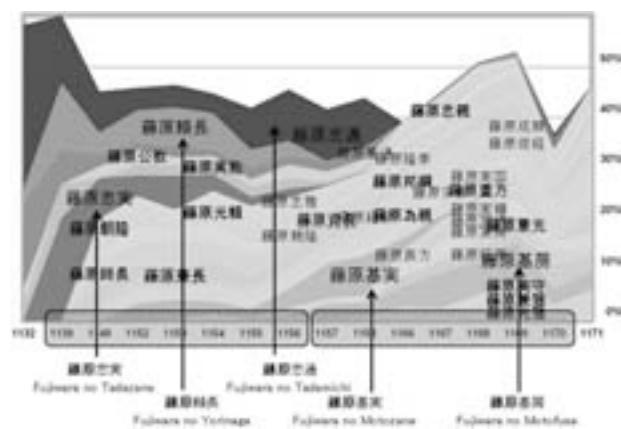


Fig. 4 Trends of family name 藤原 (Fujiwara) co-occurring over the same periods, View 2 from the repository in Fig. 3

References

- Bonnet, J.** (2004). 'New Technologies, New Formalisms for Historians: The 3D Virtual Buildings Project'. *Literary and Linguistic Computing*. Vol. 19, No. 3: 289-301.
- Heer, J. Mackinlay, J.D., Stolte, C. and Agrawala, M.** (2008). 'Graphical Histories for Visualization: Supporting Analysis, Communication,

and Evaluation'. *IEEE Transactions on Visualization and Computer Graphics*. Vol. 14, No. 6: 1189-1196.

Heer, J., Viegas, F. B., and Wattenberg, M. (2009). 'Voyagers and Voyeurs: Supporting Asynchronous Collaborative Visualization'. *Communications of the ACM*. Vol. 52, No. 2: 87-97.

James T. and Cook, K. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization Center.

Nielsen J. (1992). 'Finding Usability Problems Through Heuristic Evaluation'. *CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems*. Pp. 373-380.

Toledo, A., Thawonmas, R., Maeda, A. and Kimura, F. (2009). 'Interactive Visual Analysis of Personal Names in Japanese Historical Diary'. *DH2009*. Maryland, USA, pp. 278-280.

Van Dalen-Oskam, K. and Van Zundert, J. (2004). 'Modeling Features of Characters: Some Digital Ways to Look at Names in Literary Texts'. *Literary and Linguistic Computing*. Vol. 19, No. 3: 289-301.

Wattenberg, M. and Kriss, J. (2005). 'Designing for Social Data Analysis'. *IEEE Transaction on Visualization and Computer Graphics*. Vol. 12 no. 4: 549-557.

Weizhong, Z. and Chen, C. (2007). 'Storylines: Visual Exploration and Analysis in Latent Semantic Spaces'. *Computer and Graphics*. Vol. 31, Issue 3: 338-349.

Notes

- Stacked Graph URL: <http://www.ice.ci.ritsumei.ac.jp/~alex/infovis/hyohanki>

"Quivering Web of Living Thought": Mapping the Conceptual Networks of Swinburne's *Songs of the Springtides*

Walsh, John A.

jawalsh@indiana.edu
Indiana University

Foong, Pin Sym

Indiana University

Anand, Kshitiz

Indiana University

Ramesh, Vignesh

Indiana University

Our paper will discuss conceptual networks present in Victorian poet Algernon Charles Swinburne's mid-career collection *Songs of the Springtides* (1880) and how those networks may be represented in TEI P5 XML markup and graphic visualizations driven by the encoded text.

Swinburne's work is full of familiar signposts and nodes, such as his trademark binary oppositions and pairings: pain/pleasure, life/death, love/hate, hope/fear, sleep/death. An incredibly learned poet with an extensive range of form and allusion, Swinburne's poems are packed with often obscure references to the Bible, classical mythology, and Arthurian legend. He wrote a number of political poems addressing contemporary events. He wrote parodies of other contemporary poets, including Tennyson, Browning, and Rossetti. And as Jerome McGann has noted, "No English poet has composed more elegies than Swinburne" (McGann 293). These binary oppositions; the many biblical, mythical and legendary references; the historical and contemporary figures who are eulogized in the elegies and praised in the many tributes and dedications; the pervasive symbols of song and the sea: these elements of Swinburne's verse all serve as familiar, easily identifiable nodes of information, laden with meaning acquired through strategic repetition and structural integration into the intellectual networks of Swinburne's work. We will examine these nodes, structures and architectonic forms in one of Swinburne's most artfully crafted and carefully designed collections, *Songs of the Springtides*.

The mid-career *Songs of the Springtides* is a particularly interesting volume in the context of

inter- and intra-textual networks. For *Songs of the Springtides*, Swinburne originally planned "a little volume containing three poems upwards of 500 lines each in length, all of them in a sense sea-studies" (Swinburne *Uncollected Letters* 2:181). The three poems are: "Thalassius," "On the Cliffs," and "The Garden of Cymodoce." To this "triad of sea-studies" Swinburne added the "Birthday Ode" to Victor Hugo. Unannounced but also present in the volume are three short poems: the fifteen-line "Dedication" to Edward John Trelawny, Swinburne's "old sea king" and a friend of Shelley's (Swinburne *Uncollected Letters* 2:181); an untitled sonnet, with the first line "Between two seas the sea-bird's wing makes halt;" and another sonnet, buried in the notes to the ode for Hugo, "On the proposed desecration of Westminster Abbey by the creation of a monument to the son of Napoleon III."

This small volume is an artful example of a deliberately fashioned and architected whole connected by complex discourse networks of key concepts that operate within, across, and beyond the individual poems. Familiarity with the poems of *Songs of the Springtides* reveals a few key concepts, figures, or images of particular import and penetration: Swinburne's pantheon of literary heroes; song and music; the natural world, especially the sea; the poet; the text.

In many cases occurrences of these concepts may be identified algorithmically. However, one cannot rely on string pattern matching to find all words and phrases related to a particular concept. In the case of *song*, for instance, automated processes may be used to identify the many clear and obvious occurrences of this concept, phrases including words such as *song*, *songs*, *sing*, *singer*, *music*, etc. However, the poems also contain phrases such as the following: "lutes and lyres of milder and mightier strings," which is obviously related to music, but less susceptible to automated identification. A combination of automated and manual markup then has been used to identify and encode words and phrases related to the concepts of interest in the texts.

This notion of the text as a self-constituted network or as a part of a larger inter-textual network is found in influential writings of the major sages of poststructuralism and postmodernism. In *S/Z*, Roland Barthes writes about the text as "an entrance into a network with a thousand entrances; to take this entrance is to aim, ultimately, not at a legal structure of norms and departures, a narrative or poetic Law, but at a perspective (of fragments, of voices from other texts, other codes), whose vanishing point is nonetheless ceaselessly pushed back, mysteriously opened; each (single) text is the very theory (and not the mere example) of this vanishing, of this difference which indefinitely returns, insubmissive. (12)"

Michel Foucault, in *The Archaeology of Knowledge* writes, "The frontiers of a book are never clear-cut: beyond the title, the first lines, and the last full stop, beyond its internal configuration and its autonomous form, it is caught up in a system of references to other books, other texts, other sentences: it is a node within a network. And this network of references is not the same in the case of a mathematical treatise, a textual commentary, a historical account, and an episode in a novel cycle; the unity of the book, even in the sense of a group of relations, cannot be regarded as identical in each case. The book is not simply the object that one holds in one's hands; and it cannot remain within the little parallelepiped that contains it: its unity is variable and relative. As soon as one questions that unity, it loses its self-evidence; it indicates itself, constructs itself, only on the basis of a complex field of discourse. (23)"

More recently, Friedrich Kittler in Discourse Networks 1800/1900, building on and synthesizing the work of Barthes, Foucault, Derrida, and others, writes about literature as an information system supported and shaped by the available technologies of discourse: "An elementary datum is the fact that literature (whatever else it may mean to readers) processes, stores, and transmits data, and that such operations in the age-old medium of the alphabet have the same technical positivity as they do in computers. (370)"

These theories of the text as constituting and constituted by information networks have obvious relevance and resonance for digital humanities scholarship, much of which is engaged in explicitly identifying, encoding, and otherwise representing the information structures in texts of all kinds.

By encoding the individual information nodes, one can generate new interfaces and mechanisms for reading and navigating the text and for visualizing the patterns and interactions of the information networks operating throughout Swinburne's volume.

The authors have been working on a specific web-based visualization to represent graphically the conceptual networks at play across a series of literary texts, in this case Swinburne's *Songs of the Springtides*, and to allow users to view and browse these networks from a distance and to zoom in and focus on local clusters and concentrations of the conceptual nodes.

Our presentation will include a detailed discussion of *Songs of the Springtides* as a carefully designed information system, supported by a framework of internal and external discourse networks. Following this more theoretical discussion of Swinburne's volume, we will review and illustrate the TEI P5 mechanisms used to encode the networks and demonstrate the web-based visualization of the text.

References

- Barthes, Roland** (1974). *S/Z*. Miller, Richard (ed.). New York: Hill and Wang.
- Foucault, Michel** (1972). *The Archaeology of Knowledge and the Discourse on Language*. New York: Pantheon.
- Kittler, Friedrich A.** (1990). *Discourse Networks 1800 / 1900*. Cullens, Chris, Metteer, Michael (eds.). Stanford: Stanford University Press.
- McGann, Jerome** (1972). *Swinburne: An Experiment in Criticism*. Chicago: University of Chicago Press.
- Swinburne, Algernon Charles** (2004). *Uncollected Letters of Algernon Charles Swinburne*. Meyers, Terry L. (ed.). London: Pickering & Chatto.
- Swinburne, Algernon Charles** (1880). *Songs of the Springtides*. London: Chatto.

“It’s Volatile”: Standards-Based Research & Research-Based Standards Development

Walsh, John A.

jawalsh@indiana.edu
Indiana University

Hooper, Wally

Indiana University

You even have
my field guide. It's you I love.
I have believed so long
in the magic of names and poems
I hadn't thought them bodiless
at all. Tall Buttercup. Wild Vetch.
Often I am permitted to return
to a meadow." It all seemed real to me
last week. Words. You are the body
of my world, root and flower, the
brightness and surprise of birds.
I miss you, love. Tell Leif
you're the names of things.
—Robert Hass, “Letter”

*It's volatile because anciently painted
with wings in this manner whence came
this character ♀ for mercury.*

— Sir Isaac Newton, “Praxis,”
Babson Collection (Burndy Library Collection)
MS. 420, Huntington Library

Digital humanities scholarship often integrates humanities scholarship (literary studies, historical studies, and so on) with technological research and development. Some of this technological work takes the form of standards development. The most noteworthy example of such standards development in the digital humanities community is the Text Encoding Initiative (TEI). The TEI provides Guidelines for encoding texts for scholarly and general use. The TEI is pervasive in digital humanities and digital library contexts. It is a de facto standard developed and evolved over the past twenty some years through the efforts of a number of dedicated scholars, librarians, and technologists, and with input from the larger community of TEI users.

Another standard of significance to the digital humanities community is Unicode. Our paper presents a case-study of a successful effort to have included in the Unicode standard dozens of characters required by the *Chymistry of Isaac*

Newton, an ongoing digital humanities project to digitize and edit, study and analyze the alchemical works of Isaac Newton and to develop various scholarly tools around the collection. Unicode has become the universal character encoding standard. Unicode is nothing more, as it is certainly nothing less, than a massive mapping of characters to numbers, a mapping that seeks to accommodate all the world's languages and writing systems, including symbols of all sorts—mathematical symbols and operators, astronomical and astrological symbols, Zapf Dingbats, and many more. Operating systems, and the applications built upon them—databases, word processors and text editors, browsers, graphics software, and games—depend on such mappings, or encodings, to reliably reference, store, input, output, and display textual data. The Unicode Consortium's "What is Unicode" page <http://unicode.org/standard/WhatIsUnicode.html> accurately reports the standard's significance: "Unicode is required by modern standards such as XML, Java, ECMAScript (JavaScript), LDAP, CORBA 3.0, WML, etc., and is the official way to implement ISO/IEC 10646. It is supported in many operating systems, all modern browsers, and many other products. The emergence of the Unicode Standard, and the availability of tools supporting it, are among the most significant recent global software technology trends."

In spite of Unicode's impressive comprehensiveness, it does not include every character ever used. It does not at present, for instance, include many of the alchemical symbols found in Isaac Newton's alchemical writings. Unicode provides a "private use area," a series of reserved *code points* (the numbers assigned to characters) for projects and products to use "privately" for mapping to characters not represented in Unicode. A project like the *Chymistry of Isaac Newton* can make use of this private use area to map to characters that are not already described in the standard. A pitfall of the Private Use Area is that it is meant to be used privately; it is not suitable for easily interchangeable or interoperable data. One project's implementation of the Private Use Area could conflict with another project's. And fonts would not typically include characters for Private Use Area code points, since by their nature these codepoints are not assigned permanently to any one character but are perpetually open for *private* assignment, not as part of the public standard.

So when a project stumbles upon a rich collection of important characters and symbols that are relevant and useful beyond the interior confines of one's own project, one can make a significant scholarly contribution by documenting and describing these characters and proposing them for inclusion in the Unicode encoding standard. The alchemical symbols so common in Isaac Newton's

chymical manuscripts, are common also throughout manuscript and print alchemical literature. The graphically and semantically rich symbols also have potential utility in design, computer art, and even gaming applications. Even the few symbols that are potentially unique to Newton are worthy of consideration in the Unicode standard, given Newton's stature as one of the giants of science and the vast wealth of scientific, historical, biographical, and popular literature related to Newton.

Figure 1. Basil Valentine. "A Table of Chymicall & Philosophicall Charecters with their signs." *The Last Will and Testament of Basil Valentine*, 1671. These and other symbols are commonly found in Newton.

The process by which one moves a Unicode proposal through the development, review, and approval process is formal and rigorous. It is very rewarding in fostering a better understanding of one's source material and in pointing the way to undiscovered or avoided basic research questions. To encode and identify characters and symbols, one must name the things, and naming is indeed a very difficult and powerful task, a task often challenged and enriched by puzzling ambiguity and obscurity. The process is very rewarding also because it is very much peer-

reviewed. Our proposal greatly benefited from an iterative review and excellent advice, challenging questions, and constructive criticism from a number of very smart, helpful, interested experts serving on the Unicode Technical Committee (UTC).

Our paper provides a case-study of one project's navigation through the Unicode proposal, review, and approval process. We also provide a more theoretical discussion, illustration, and examination of the mutually beneficial relationship between technical standards development and basic humanities research.

References

- Unicode Consortium** (15 June 2009). *What is Unicode?*. <http://unicode.org/standards/whatIsUnicode.html> (accessed 15 Nov 2009).
- Newman, William R. (ed.)** (9 May 2008). *The Chymistry of Isaac Newton*. <http://www.chymistry.org/> (accessed 15 Nov 2009).

Quelques réflexions sur l'effet propédeutique des catalogues des collections des musées en ligne

Welger-Barboza, Corinne

corinne.welger@wanadoo.fr

Université Paris 1 Panthéon-Sorbonne Institut National d'Histoire de l'Art (INHA) Observatoire critique des ressources numériques pour l'histoire de l'art et l'archéologie, France

Pour la discipline de l'histoire de l'art, l'environnement du Web a pu sembler propice à un rééquilibrage ou à de nouvelles alliances entre institutions patrimoniales et académiques, ne serait-ce que pour instaurer «un nouveau partage de l'image» (Welger-Barboza, 2006). L'accessibilité croissante de corpus d'images et de sources (Greenhalgh, 2004) est restée majoritairement le fait des musées. Et l'on peut considérer que – à l'exception notable des antiquisants et des médiévistes, équipes souvent pluridisciplinaires – les enseignants-chercheurs en histoire de l'art sont encore faiblement à l'initiative de corpus d'étude outillés (*Archives*) répondant aux démarches et méthodes propres à l'histoire de l'art (CLIR Seminar Report 1988).

L'implication des musées en ligne depuis 15 ans jusqu'au Web 2.0 a donné lieu à une littérature abondante.¹ Mais l'attention portée aux publics par les protagonistes inclut rarement les historiens de l'art, c'est-à-dire la visée de l'exploitation scientifique de ces productions, tandis qu'il est fait peu de cas des catalogues des collections, en tant que tels. Pourtant, l'offre en ligne des musées mérite d'être prise en considération du point de vue de l'histoire de l'art, émanant de l'un de ses deux corps (*Haxthausen*, 2002). Dans cette perspective, nous voulons montrer comment les musées prennent une part non négligeable au développement des *Digital Humanities* et plus particulièrement, au titre de la mise en ligne des catalogues de leurs collections.

Au croisement des missions professionnelles, scientifiques, éducatives du musée et de son immersion dans l'évolution de l'environnement sociotechnique du web, la présentation des collections des musées en ligne s'inscrit dans le prolongement d'un outil stratégique de la discipline, le catalogue. En tant que tel, il est tout à la fois l'héritier de l'histoire du catalogue imprimé dont le geste initial revient au Catalogue Crozat (Recht, 1996), des tables d'inventaires des collections

et de l'intelligence mécanicienne du «montrer-s'orienter-classer» dont Patricia Falguières situe la source dans les Chambres des merveilles (Falguières, 2003; 1996). Informatisés, les corpus instrumentés que sont devenus les catalogues explorent des relations nouvelles entre les textes et les images, entre l'œuvre singulière et l'ensemble formé par la collection. Grâce à leur accessibilité en ligne, ces catalogues, directement issus de l'activité des conservateurs, exposent le socle commun aux «deux histoires de l'art»; en même temps, ils font partie intégrante de l'offre éducative du musée à l'adresse de publics indéterminés.

A partir de nos recherches actuelles sur les bases des collections des musées en ligne, nous étudions plus particulièrement les choix opérés pour l'indexation ainsi que les approches des œuvres par l'image. La problématique générale tend à identifier et caractériser, à partir de la pluralité des propositions, les rapports établis entre, d'une part, les opérateurs de navigation au sein des corpus et, d'autre part, l'offre d'un espace d'examen des objets individués.

1. Entre décrire et voir, l'ajustement du point de vue

A ce stade, nous envisageons que ces deux composantes essentielles, à savoir: l'indexation, plus particulièrement l'indexation sujets, et les outils de visualisation-manipulation de l'image – d'ailleurs également caractéristiques des corpus d'étude outillés (*Archives*) – participent ensemble à la mise en œuvre de points de vue sur les œuvres et la collection. Cette question du point de vue, prise dans sa polysémie, unit les deux familles d'outils mentionnés ci-dessus. L'indexation, outre l'identification, désigne, oriente la recherche et constitue un dispositif de médiation entre l'autorité de l'institution et les utilisateurs de la base de données. L'instauration d'un espace virtuel d'étude des objets sélectionnés les dégage du référent du catalogue imprimé; l'image est ainsi disponible au traitement et livrée à une série de gestes de la part des utilisateurs.

La communication illustrera une étape du travail en cours par deux exemples: le système d'indexation adopté par *Tate Online* et les instruments de vision rapprochée et de comparaison du LACMA (*Los Angeles County Museum of Art*). Ces deux propositions, outre leurs qualités intrinsèques, sont exemplaires dans la mesure où elles offrent une représentation plausible et plus achevée que d'autres² des fondamentaux revendiqués unanimement par la discipline, à savoir la description et l'examen visuel des œuvres.

Les descripteurs de *Tate Online* (Glossary³) sont issus pour une grande part de l'ICONCLASS, vocabulaire contrôlé à l'usage des professionnels auquel des catégories thématiques ont été ajoutées. Ainsi, le lexique de la *Tate Online* satisfait à l'exercice académique de description, dans la mesure où les catégories retenues et leurs ramifications s'appliquent à déterminer le rattachement aux courants artistiques, à identifier les sources et l'iconographie tandis qu'une forme d'interprétation prend appui sur des «concepts, sentiments, idées». En renfort, les thèmes (nature, occupations, loisirs, etc.) établissent un pont entre des approches interprétatives spécialisées et non spécialisées. Cet appareil discursif permet la lecture et la situation de chaque œuvre au sein de la collection; en outre, choix original, il s'actualise en s'exposant lors de l'affichage de chaque œuvre. Enfin, des rebonds sont proposés à l'utilisateur, grâce au croisement de termes placés à des degrés différents d'arborescence dans chaque catégorie. La formule laisse du jeu à l'utilisateur pour circuler au sein du cadre construit par le musée qui propose une lecture informée et savante partie prenante d'une propédeutique de l'histoire, du point de vue du musée. Un véritable complexe composé des descripteurs, de l'exposition de leur architecture et des rebonds de classification soutient la stratégie de médiation du musée.

Le «Viewer» du LACMA présente, d'une autre façon, un cas de figure intéressant. Dégagées du catalogue, les œuvres sélectionnées s'inscrivent dans un espace dédié à la visualisation. Elles se prêtent à la possibilité d'une vision exceptionnellement rapprochée grâce à la numérisation de bonne résolution d'une partie de la collection (malgré des inégalités de traitement). Le point de vue sur l'objet s'est autonomisé des catégories pré-établies pour se donner libre cours. Si le LACMA n'est pas le seul à proposer des visions rapprochées d'un degré important, il présente l'intérêt original de combiner le zoom à un comparateur d'images, dressant ainsi le cadre virtuel d'une table lumineuse. Ici, le dispositif invite à un exercice propédeutique: l'approche méthodique de l'examen visuel et comparatif. Incontestable quant à sa pertinence dans la pratique disciplinaire, cette représentation suscite néanmoins l'interrogation à un double titre.

Hypostasiée, l'approche visuelle ne réduit-elle pas l'ensemble des méthodes, documentaires notamment, dans lesquelles elle s'inscrit pour faire sens, y compris dans une perspective «attributionniste»? En second lieu, la popularisation de la posture de l'examen visuel – auquel le LACMA contribue – laisse dans l'ombre la diversité des emplois de la vision rapprochée dans l'étude des œuvres, grâce à l'imagerie numérique, selon des conditions variables: vision des détails ? Vision

de l'invisible à l'œil nu ? Ainsi de la relation ambivalente entre l'œuvre et le document : vision rapprochée du document ? Vision rapprochée de l'œuvre par le document, dans la lignée de l'imagerie scientifique ? En ce sens, n'avons-nous pas affaire à une représentation mimétique de gestes emblématiques de l'historien de l'art dans son cabinet d'étude ?

Ainsi, les musées contribuent d'une façon décisive à la prégnance progressive des *Digital Humanities* dans la représentation et l'éducation de l'histoire de l'art à l'adresse d'un large public. Les catalogues des œuvres en ligne proposent des figures plurielles où se manifeste l'abandon du référent imprimé au profit de jeux dynamiques où s'affirme l'autonomie des dimensions discursives et visuelles. Celles-ci sont porteuses d'enrichissements et de contradictions. D'autres approches, portées par des collections, des corpus formés et informés dans le cadre de l'université - notamment des collections individuelles ou collectives – peuvent bénéficier de ces propositions dont le musée est prolixie; de ce point de vue, on peut leur attribuer encore un effet propédeutique, quant à l'exploitation de la culture numérique dans le cadre de la discipline. En effet, ces réalisations offrent non seulement des modèles en termes d'instrumentation mais encore elles favorisent la réflexivité. Notamment, leur analyse favorise l'introspection de la discipline sur ses méthodes; c'est le propre de la formalisation drainée par l'instrumentation numérique, dans toutes les disciplines.



Figure 1: Paolo and Francesca da Rimini, Dante Gabriel Rossetti (1828-1882), Subjects Page –Tate Online Collection

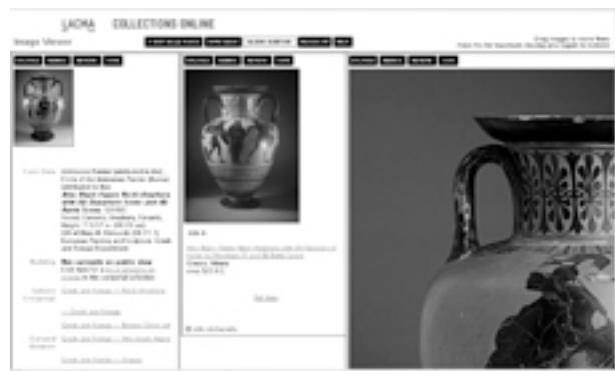


Figure 2: Attic Black – Figure Neck – Amphoras (Greece circa – 520) – Image Viewer – LACMA Collection Online

References

Art Historians and Their Use of Illustrated Texts: Scholarly Resources In Art History, Issues in Preservation – Report of the Seminar, Spring Hill, Wayzata, Minnesota, September 29 - October 1, 1988. <http://www.clir.org/pubs/reports/cpaarth/cpaarth.html> (accessed 30/10/09).

Falguières, P. (2003). 'Les chambres des merveilles'. *Cahiers du musée national d'art moderne*. **nº 56/57**: 5-20.

Greenhalgh, P. (2004). 'Art History'. *A Companion to Digital Humanities*. Susan Schreibman, Ray Siemens, John Unsworth (ed.). Oxford: Blackwell. <http://www.digitalhumanities.org/companion>.

Haxthausen, C.W. (ed.) (2002). *The Two Art Histories – The Museum and the University*. Williamstown: Sterling and Francine Clark Institute.

Recht, R. (1996). 'La Mise en ordre, Note sur l'histoire du catalogue'. *Cahiers du musée national d'art moderne*. **nº 56/57**: 21-35.

Welger-Barboza, C. (2006). 'Vers un nouveau partage de l'image'. *Observatoire critique des ressources numériques pour l'histoire de l'art et l'archéologie*. http://www.observatoirecritique.org/article.php3?id_article=26 .

Notes

1. Le corpus des actes des conférences de Museum and The Web et de ICHIM en témoigne. [En ligne] <http://www.archimuse.com/publishing/papers.html>. Voir aussi le blog Museum 2.0 <http://museumtwo.blogspot.com>
2. Le corpus d'étude comprend principalement les bases des collections des grands musées internationaux : Musée National d'Art Moderne, Rijksmuseum, Tate online, V&Amuseum, National Gallery of Art (London), National Galery of Art (Washington D.C.), Metropolitan Museum of Art, Museum Of Modern Art, San Francisco Museum Of Modern Art, Los Angeles County Museum Of Art, Art Institute of Chicago, Indianapolis Museum of Art, etc.

3. <http://www.tate.org.uk/servlet/SubjectSearch>

"Any more Bids?": Automatic Processing and Segmentation of Auction Catalogs

West, Kris

Kris.West@gmail.com
JSTOR, USA

Llewellyn, Clare

Clare.Llewellyn@ithaka.org
JSTOR, USA

Burns, John

John.Burns@ithaka.org
JSTOR, USA

This paper details work that has been conducted through a collaborative project between JSTOR, the Frick Collection and the Metropolitan Museum of Art. This work, funded by the Andrew W. Mellon Foundation, was to understand how auction catalogs can be best preserved for the long term and made most easily accessible for scholarly use. Auction catalogs are vital for provenance research as well as for the study of art markets and the history of collecting. Initially a set of 1604 auction catalogs, over 100,000 catalog pages, was digitised – these catalogs date from the 18th through the early 20th century.

An auction catalog is a structured set of records describing items or lots offered for sale at an auction. The lots are grouped into sections – such as works by a particular artist, each of the sections are then grouped into a particular sale – this is the actual event that happened in the sale room, and then these sales are grouped together in the auction catalog. The auction catalog document also generally includes handwritten marginalia added to record other details about the actual transaction such as the sale price and the buyer.

A repository was constructed – this holds and provides access to page images, optical character recognition (OCR) text and database records from the digitised auction catalogs. In addition a website was created that provides public access to the catalogs and automatically generated links to other collections. This site offers the ability to search and browse the collection and allows users to add their own content to the site.

When searching a user may only be interested in a single item within a larger catalog, therefore, to facilitate searching the logical structure of the

catalog needs to be determined in order to segment the catalog into items. The catalogs are extremely variable in structure, format and language, and there are no standard rules that can divide the catalog into the lots, sections and sales. Therefore, machine-learning techniques are used to generate the segmentation rules from a number of catalogs that have been marked up by hand. These rules are then applied to classify and segment the remaining catalogs.

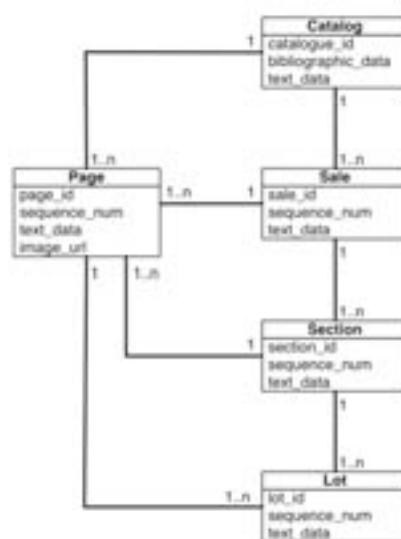
The focus of this paper is the research and creation of a system to automatically process digitised auction catalog documents, with the aim to automatically segment and label entities within and create a logical structure for each document. The catalogs processed are in an XML format produced from physical documents via an OCR process. The segmentation and assignment of entity types will facilitate, deep searching, browsing, annotation and manipulation activities over the collection. The ability to automatically label previously unseen documents will enable the production of other large scale collections where the hand labelling of the collection content is highly expensive or unfeasible.

The catalog, sale, section, lot model requires that the content of the document be distributed between these entities, which are themselves distributed over the pages of the document. Each line of text is assigned to a single entity, whole entities may be contained within other entities (a logical hierarchy), and a parent entity may generate content both before and after its child entities in the text sequence. This hierarchical organisation differentiates the problem of automatically labelling auction catalog document content from other semantic labelling tasks such as Part of Speech labelling (Lafferty et al., 2001) or Named Entity Recognition (McCallum and Li, 2003). In these tasks the classes or states can be thought of as siblings in the text sequence, rather than as having hierarchical relationships. Hence, the digitisation of auction catalog documents may require a different set of procedures to that applied to, for example, the digitisation of Magazine collections (Yacoub et al., 2005) or scholarly articles (Lawrence et al., 1999).

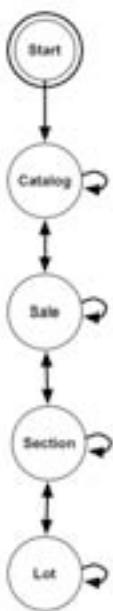
Although a particular document model is assumed throughout this work, the theory and tools detailed can be applied to arbitrary document models that incorporate hierarchical organisation of entities.

Techniques that are successfully applied to other Natural Language Processing and document digitisation tasks may be applied to this problem. Specifically, we have developed task appropriate feature extraction and normalisation procedures to produce parameterisations of catalog document content suitable for use with statistical modelling techniques. The statistical modelling technique

applied to these features, Conditional Random Fields (CRFs) (Sutton and McCallum, 2007), models the dependence structure between the different states (which relate to the logical entities in the document) graphically. A diagrammatic representation of the auction catalogue document model is given in figure 1a and an example of a model topology that might be derived from it is given in figure 1b. It should be noted that CRFs are discriminative models, rather than generative models like HMMs, a property that may be advantageous when such models are applied to NLP tasks (Lafferty et al., 2001).



(a) Document data



(b) FST transition grammar

Figure 1: Document model for an auction catalog and a Simple Finite State Transducer topology derived from it

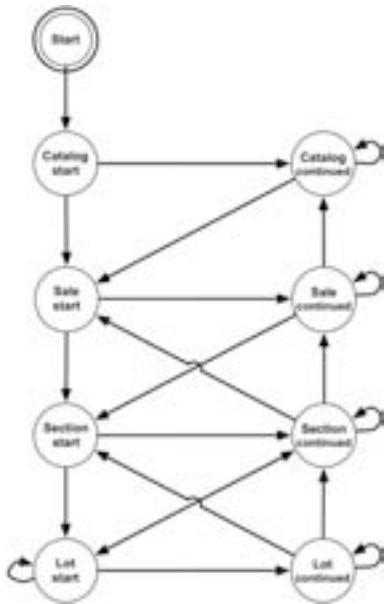


Figure 2: FST transition grammar extended to incorporate start and continue states for each entity type

The application of such techniques to hierarchically structured documents requires the logical structure of a document to be recoverable from a simple sequence of state labels. The basic transition grammar shown in figure 1b is not appropriate for this task as it is impossible to differentiate concurrent lines of a single entity from concurrent lines from two entities of the same type and relationships between a single ‘parent’ entity and multiple ‘child’ entities. These issues may be addressed by using two states, a start and a continuation state, in the model, to

represent content relating to each entity, as shown in figure 2. This modification allows the full logical structure to be recovered by post-processing the sequence of state labels.

In order to train any automated statistical learning procedure, a suitable sample of the data set must be labeled, usually by hand, to provide a target for learning or a ground-truth for evaluation. Hand labelling an XML-based representation of a document, produced via an OCR process, can be a difficult and frustrating task. To facilitate the fast labelling of documents, and thereby maximise the quantity of ground-truth data that could be produced, a cross-platform document segmentation user application was developed, shown in figure 3, that is able to format the OCR data for display and allow the user to simply and rapidly mark-up each document. This application could easily be adapted for use with other data sets and is a useful output of this work that can be used independently.

Initial experiments with the segmentation and labelling system were conducted on a hand labeled collection of 266 auction catalogs, comprising just less than 400,000 lines of text content. A detailed breakdown of the content of these documents

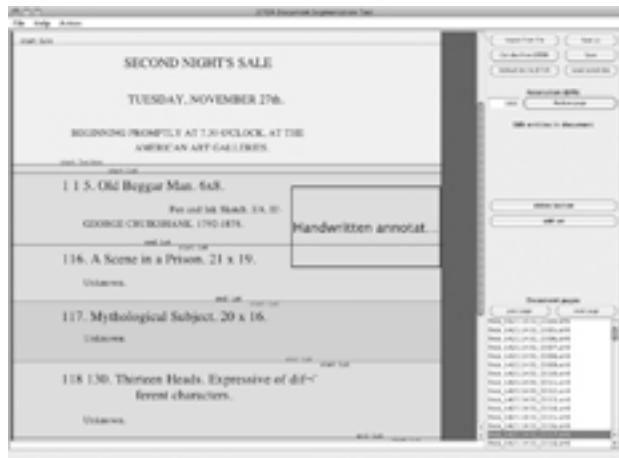


Figure 3: Document segmentation user application

is provided in table 1. These experiments compare a number of variants of the system and analyses are provided to aid in the selection of parameters for the system. The experiments are evaluated using both the retrieval statistic F-measure (Baeza-Yates and Ribeiro-Neto, 1999) and an adaption of the WindowDiff metric (Pevzner and Hearst, 2002). The system has been found to produce a satisfactory level of performance to facilitate the automated processing of auction catalog content. A breakdown of the results achieved by the system is given in table 2.

Entity	Number of lines
start-Catalog	266
cont-Catalog	41,275
Catalog	41,541
start-Sale	556
cont-Sale	3,469
Sale	4,025
start-Section	4,109
cont-Section	19,415
Section	23,524
start-Lot	91,240
cont-Lot	229,051
Lot	320,291
Total	389,381

Table 1: Number of lines corresponding to each entity type

Analysis of the errors produced by the system show that, owing to the hierarchical nature of the document model, a small number of errors at critical points in the content can lead to a large number of subsequent, concurrent errors. Unfortunately, these critical points in the ground-truth sequences are represented by the smallest quantities of content in the collection and therefore may form the weakest parts of the model. However, this also means that a small number of corrections at these key points could enable the correction of a much larger number of errors in the output. This is a useful property, as one of the design goals of this system is to facilitate the integration of feedback from users of the digitised documents into the statistical model and thereby allow the segmentations produced to improve over time.

The feedback is used to improve the model by preserving any confirmations or corrections, made by users, of the segmentation output from CRF model. This preservation is achieved by reapplying the model to the document, which has been partially labeled by users, and forcing it to pass through the user indicated states as it determines a new sequence of state labels. This allows a user to supply corrections of major segmentation errors, such as a missing high-level entity (e.g. a sale), or minor errors, such as a single mislabeled line belonging to a lot. A user supplied correction to a major segmentation error could correct the labelling of many lines, for example by opening the Sale at the correct point and allowing the model to estimate weights for Sections and Lots thereafter, whereas minor corrections may be simply preserved so that they don't need to be reapplied to the re-segmented document.

Metric	Score
WindowDiff	0.103
F1 cont-Catalog	0.709
F1 start-Lot	0.870
F1 cont-Lot	0.934
F1 start-Sale	0.444
F1 cont-Sale	0.347
F1 start-Section	0.483
F1 cont-Section	0.548

Table 2: Results achieved by the CRF-based system calculated over 5-fold cross-validation of the hand-labelled dataset

References

- Baeza-Yates, R., Ribeiro-Neto, B.** (1999). *Modern Information Retrieval*. Addison-Wesley Publishing Company.
- Lafferty, J., McCallum, A., Pereira, F.** (2001). 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data'. *Machine Learning-International Workshop then Conference*. Citeseer, pp. 282–289.
- Lawrence, S., Giles, C., Bollacker, K.** (1999). 'Digital libraries and autonomous citation indexing'. *IEEE computer*. **32(6)**: 67–71.
- McCallum, A., Li, W.** (2003). 'Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons'. *Seventh Conference on Natural Language Learning (CoNLL)*.
- Pevzner, L., Hearst, M.** (2002). 'A critique and improvement of an evaluation metric for text segmentation'. *Computational Linguistics*. **28(1)**: 19–36.
- Sutton, C., McCallum, A.** (2007). *An Introduction to Conditional Random Fields for Relational Learning. Introduction to statistical relational learning*. Pp. 93.
- Yacoub, S., Burns, J., Faraboschi, P., Ortega, D., Peiro, J., Saxena, V.** (2005). 'Document digitization lifecycle for complex magazine collection'. *Proceedings of the 2005 ACM symposium on Document engineering*. ACM New York, NY, USA, pp. 197–206.

Mandoku – An Incubator for Premodern Chinese Texts – or How to Get the Text We Want: An Inquiry into the Ideal Workflow

Wittern, Christian

cwittern@gmail.com

Kyoto University

Premodern Chinese texts pose problems that are difficult to accommodate with the current TEI text model, which bases the main hierarchy of a text on its structural content, rather than on a hierarchy that models the pages, lines and character positions. For the TEI, this is a sensible decision and has led to the abolishment of elements like `<page>` and `<line>` in the latest release of the Guidelines. For premodern Chinese texts however, especially texts that are transmitted as manuscripts or woodblock printings and have not yet seen a modern edition printed with movable type (let alone as, more recently, computerized typesetting), establishing the structural hierarchy of the text content is, together with the even more daunting question of establishing the proper characters of the text (on which see below), an important part of the research question that motivates the digitization of the text. Requiring an answer to this question before a proper electronic text can be created makes this intractable in the digital medium and glosses over an important leap of faith in the creation of a TEI encoded text. In this paper, I will try to trace some of the implications and propose an approach that allows different models of the text for different stages in the encoding process, thus closer modeling the process of the creation of an electronic text.

To arrive at a text properly encoded according to the TEI Guidelines is not a straightforward process, but in the setup described here requires a detour through at least three stages:

- Draft input of the text without any further markings;
- An incubator phase, in which the text is dealt with as a series of pages (or scrolls), lines and characters;
- The mature text, based on the structural model of a TEI `<text>`, which is available for further refinement;

Of these steps, the second one is at the center of attention in this paper, which will include the discussion of the following three aspects:

1. A text model according to these requirements;
2. Mandoku, an application that allows manipulating the text;
3. A transform that specifies how a text conforming to this specification can be turned into a TEI encoded text.

1. Different Models for the Same Text

The structural, content based hierarchy of the text has to be established as part of the research process. For this reason, the text at this stage uses the only hierarchy available, that is the one that is based on how the text is physically recorded on the writing surface in the edition used. During the process of working with the text, milestone-like elements are inserted at the starting points of elements of interest, using the incubator as described in the next section. Headings are numbered according to their nesting depth as in a HTML document; this forms the base for their transformation into regular TEI nested `<div>`s followed by `<head>` elements.

2. The Incubator: Mandoku

The tool used to manipulate a premodern Chinese text in the incubator phase has been called Mandoku. It makes it possible to display a digital facsimile of one or more editions and a transcribed text of these editions side by side on the same screen. From there, the texts can be proofread, compared and annotated. A special feature is the possibility to associate characters of the transcription with images cut from the text and a database of character properties and variants, which can be maintained while operating on the text. Interactive commands exist also to assist in identifying and record structural and semantic features of the texts.

One of the major obstacles to digitization of premodern Chinese texts is the use of character forms that are much more ideosyncratic than today's standard forms. Since in most cases they cannot be represented, they are exchanged during input for the standard forms. This is a potentially hazardous and error-prone process at best, and completely distorts the text in worse cases. To improve on this situation and to make the substitution process itself available to analysis, Mandoku uses the position of a character in a text as the main means of addressing, allowing the character encoding to become part of the modelling process, thus making it available to study and analysis, which in turn should make the process

of encoding more tractable even for premodern texts. The current model is still experimental, but initial results have been encouraging.

Mandoku is work in progress and is developed as part of the Daozang jiyao project at the Institute for Research in Humanities, Kyoto University by Christian Wittern. In this paper, an emphasis will be placed on the different models of a text that are underlying the different stages of preparation of a text and the friction, but also benefits, that arise out of such a situation. The following is a screenshot of the main interface, displaying a facsimile and a transcribed version of the same text side by side.



Fig. 1. *Mandoku* in action

3. Transform to TEI <text>

Finally, as a proof of concept, a XSLT script has been developed that performs an algorithmic transformation from the text in the intermediate format to a text as it has to appear as content of the TEI <text> element.

This produces a new version of the text with an inverted hierarchy: The primary hierarchy now is the content hierarchy, whereas the hierarchy of the text bearing medium is demoted to a secondary one, represented by milestones. None of these hierarchies is a priori superior to the other, but in the context of the Daozang jiyao project the purpose of preparing the texts is to make it available for a study of the collection, so the emphasis during the later stages in the life of the text will lie on the content hierarchy. The problem of overlapping hierarchies, which is such a scratching itch for many text projects, poses itself thus in a slightly different incarnation: The different hierarchies occur in two different stages of preparation of the text, which require different viewpoints, but not simultaneous presentation, which makes it easier to accommodate the two in our workflow.

The preparation of a TEI encoded representation of the texts is however not the ultimate goal of the project. The next phase requires analytical interaction with the text for which again the TEI representation might not be the ideal format to

work with, so there might be a number of different, purpose-specific derivative formats generated from the TEI texts. They will maintain the required information to refer additional information back to the master files kept in TEI, and to be able to participate in the ongoing evolving of the master text, to which transcriptions of more witnesses will be added, but will otherwise also contain additional commentary, translation and other information that will not belong to the original file. The details of this part of the system are under consideration now and will be the topic for another presentation.

4. Conclusions

The current TEI text model does not allow the direct description of the document as it appears on a text bearing surface without also establishing a content hierarchy. For this reason, a temporary encoding strategy had to be developed, which is TEI conformant to the letter, but not to the spirit of the TEI Guidelines by wrapping all of the text content in one giant <p> (or possibly <ab>) element. Only after the structural hierarchy has been established is it possible to make a transformation to a truly conformant and satisfying TEI document. The slight feeling of uneasiness that this workaround causes might go away once the new <document> element proposed by the TEI working group on genetic editions has been adapted to the Guidelines and can be used for the phase of work in the incubator, thus making the text fully TEI conformant from the beginning. On the other hand, this project also clearly demonstrates the necessity of being able to represent the document in its own right in a TEI text, even if in the context of this project the documentary part is considered transitory.

References

- Esposito, Monica** (2009). 'The Daozang Jiyao Project: Mutations of a Canon'. *Daoism: Religion, History and Society*. 1: 95-153.
- TEI Consortium (ed.)**. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 1.6.0. <http://www.tei-c.org/Guidelines/P5/> (accessed 2010-03-09).
- Wittern, Christian** (2007). 'Patterns of Variation: The textual Sources of the Chinese Buddhist Canon as Seen through the CBETA edition'. *Essays on East Asian Religion and Culture*. Christian Wittern and Shi Lishan (ed.). Kyoto: Editorial committee for the Festschrift in honour of Nishiawaki Tsuneki, pp. 209-232. <http://kanji.zinbun.kyoto-u.ac.jp/~wittern/data/nw-fs/fs-wittern.pdf> (accessed 2010-06-05).

Wittern, Christian (2009). 'Digital Editions of premodern Chinese texts: Methods and Problems – exemplified using the Daozang jiyao'. *Early Chán Manuscripts among the Dūnhuáng Findings – Resources in the Mark-up and Digitization of Historical Texts*. University of Oslo, Sep. 28 to Oct. 3, 2009, preprint PDF available. <http://kanji.zinbun.kyoto-u.ac.jp/~wittern/data/digital-editions-dzjy.pdf> (accessed 2010-03-09).

Towards a Computational Narration of Inner World

Zhu, Jichen

jzh@mail.ucf.edu

Department of Digital Media, University of Central Florida

Narrative, as it evolves with technological developments, constantly reinvents itself in order to better capture new social orders and individual experiences. As an emerging cultural expression, however, most computer-generated narrative works are still restricted to an action-based, goal-driven aesthetics, leaving little space for characters' inner world. This paper proposes to expand the range of computer-generated narrative by addressing this imbalance between the "physical" and the "internal." It presents our approach for algorithmically narrating characters' inner world by leveraging the synergy between modernist stream of consciousness literature and contemporary research in artificial intelligence and cognitive science. The *Riu* system, a text-based computational narrative system inspired by Woolf's novel *Mrs. Dalloway*, is provided as a case study towards this new direction.

1. Computer-Generated Narrative

Contemporary forms of narrative have evolved rapidly as digital technologies continue to be integrated in modern society. New conventions at the levels of both content and discourse have been established to reflect the constantly changing relationship between human and technology. For instance, popular science fictions of the 1980s (e.g., *The Terminator*) embodied the prevailing cyborg discourse and confusions of human identity within the Cold War context (Edwards, 1996). Similarly, hypertext fictions in the 1990s instantiated the postmodernist mentality of its time by turning everything – writer, reader, and society – into fragments (Johnson-Eilola, 1997). In this regard, the emerging form of *computer-generated narrative*,¹ that is, stories produced by computer algorithms, may offer an important cultural expression to portray our increasingly technology-dependent modern life.

Compared to other forms of electronic literature (e.g., hypertext fictions), the strict technological requirements for computer-generated narrative have confined its development largely to the computer science community, particularly artificial intelligence (AI). Over the past decades, serious attempts have been made to integrate narratology theory

into existing AI framework for story generation (Bringsjord & Ferrucci, 2000; Cavazza & Pizzi, 2006; Mateas, 2002; Meehan, 1976). Despite the considerable progress the community has made, the expressive power of computer-generated narrative is still limited compared to its non-digital antecedents. In particular, this paper is concerned with the prominent goal-driven, problem-solving aesthetics that dominate many story generation systems. A salient example is the *Tale-Spin* system (Meehan, 1976), which generates stories in the spirit of: Joe Bear was hungry; Joe couldn't reach his food because of certain obstacles; Joe resolved the issues; Joe got his food.

It is true that recent narrative systems have evolved in numerous aspects since then. Nevertheless, this ultra-rational, "behaviorist" narrative style, afforded by Meehan's now-widely-adopted planning-based framework, has remained and been taken for granted by many practitioners. As we become more aware of digital media's capability of constructing subjective mental imagery and evoking users' imagination and awareness (Harrell, 2009), it is crucial to revisit some of these early assumptions of computer-generated narrative and critically understand the expressive affordances as well as restrictions of the computational techniques that we use.

This paper proposes to expand the spectrum of computer-generated narratives by focusing on the richness of characters' inner world, hidden behind the external world of actions. This approach aligns with modernist writers' concerns of depicting "hidden life at its source" (Woolf, 1957 [1925]). Notice this is not a strong AI attempt to model human (semi-)consciousness. Instead, the goal is to explore new ways of conveying human subjectivity and life stories by algorithmically generating *narratives* that are reminiscent of similar phenomena. Informed by modernist literary techniques (particularly Virginia Woolf's work), cognitive science discoveries and AI, this paper proposes a new approach for generating inner narratives and presents initial results from our on-going narrative project *Riu*.

2. Synergy of the Old and New

As argued elsewhere (Zhu & Harrell, 2010 (forthcoming)), the overlooked synergy between stream of consciousness literature, artificial intelligence (AI), and cognitive science provides valuable insights to generating stories about characters' inner world. In their respective historical contexts, both stream of consciousness literature and AI challenged the domination of behaviorism by turning *internally* to the human psyche. Rejecting the literary representation of characters as the "external man", modernist writers such as Virginia

Woolf and James Joyce invented techniques to depict the moment-by-moment psychic existence and functioning of the "internal man" (Humphrey, 1954). Similarly, AI broke away from the behaviorism-dominated scientific community in the 1950s and legitimated human mental constructs, such as knowledge and reasoning, as crucial subjects of scientific inquiries.

The differences between AI and modernist literature further dissolve when we take account of recent cognitive science theory, a sister field of AI. Stream of consciousness literature's key concern with pre-speech level of consciousness, minimally mediated by rationality and language, is echoed by new discoveries in cognitive linguistics. Recent research (Fauconnier & Turner, 2002) has confirmed that the vast cognitive resources of "backstage cognition" are called up unconsciously when we engage in any language activity.

3. Generating Inner Narratives

Generating narratives about characters' inner world requires innovation at the story content, discourse and algorithmic levels. The techniques in stream of consciousness literature offer invaluable insights into literary representations of inner life, such as Woolf's loosely structured plot, the "caves" of characters' past (Woolf, 1957 [1925]), and various modes of interior monologues (Cohn, 1978).

The insights from planning-generated stories illustrate the impact of underlying computational techniques. Substantial changes at the algorithmic level therefore are needed to incorporate the new content and aesthetic requirements. As we have argued (Zhu & Ontañón, 2010), computational analogy, influenced by related cognitive science studies, is one of the promising directions towards our goal. Its emphasis on similarities and associations between different constructs is particularly useful to establish connections between external events and inner thoughts (e.g., the action of "buying flowers" and flower-related memories). Computational analogy may also be used to depict "the train of thoughts" by connecting a sequence of related events, one after another.

4. Case Study: *Riu*

Our generative narrative project *Riu* is an on-going attempt to computationally generated stories about characters' inner world. Inspired by Woolf's novel *Mrs. Dalloway* (Woolf, 2002 (1925)), this project harnesses computational analogy at different levels of story generation for various narrative effects (Zhu & Ontañón, 2010). Similar to our earlier conceptual-blending-based (Fauconnier & Turner,

2002) project *Memory, Reverie Machine* (Zhu & Harrell, 2010 (forthcoming)), *Riu* is explicitly geared towards algorithmically narrating characters' inner world, through the depiction of characters' unrolling thoughts and subjective variations of such thoughts based on user interaction. An excerpt of system output at the current stage of development can be found in Fig 1.

```

Walking on the street, Aliee suddenly saw a cat stopped in front of him.
He used to have a bird when he was young.
He was so fond of it that he played with it everyday.
One day the bird died, Aliee became extremely sad for weeks.
Aliee hesitated for what to do ...
(FEED, IGNORE, OR PLAY)
> Play
What if the cat also dies... be paused.
(FEED, OR IGNORE)
> Feed
Aliee took some food from his bag, gave it to the cat and walked away.

```

Figure 1. Sample output of *Riu* (including user input)

Rather than ordering events in ways that lead to a desired goal, as in many planning-based systems, *Riu* adopts the computational analogy algorithm of structural mapping (Falkenhainer, Forbus & Gentner, 1989; Gentner, 1983) to connect external events with inner thoughts. The overall plot is loosely structured with potential inner thoughts between various external events. The overall story intends to invoke what Woolf described as "no plot, no comedy, no tragedy, no love interest or catastrophe" (Woolf, 1957 [1925]) and thus focuses primarily on characters' psyche.

In Fig 1, the protagonist's memory of a pet bird (italicized) is triggered because of its similarity to his encounter of another animal in the external story world. Importantly, inner events such as memories in *Riu* also have a significant impact on the protagonist's external actions. In the case of Fig 1, when the user chooses to play with the cat on the street, the system infers that this action may, again, lead to the death of the cat and hence the protagonist's sadness. In order to prevent that from happening, the system ignores this user command and removes "play" from the next user input options.

In conclusion, computer-generated narrative is an important frontier in digital humanities as a potential cultural expression innate to our contemporary society. Drawn on the synergy between stream of consciousness literature, artificial intelligence, and cognitive science, this paper proposed a new approach for expanding its narrative spectrum and focusing on characters' inner world with its connection to external behaviors and environments. The *Riu* system was used to illustrate the initial results of our work, which suggests new expressive possibilities afforded by our approach.

References

- Bringsjord, S. & Ferrucci, D. A.** (2000). *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*. Hillsdale, NJ: Lawrence Erlbaum.
- Cavazza, M. & Pizzi, D.** (2006). 'Narratology for Interactive Storytelling: A Critical Introduction'. *Proceedings of TIDSE*. 2006, pp. 72-83.
- Edwards, P. N.** (1996). *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge, Mass.: MIT Press.
- Falkenhainer, B., Forbus, K. D. & Gentner, D.** (1989). 'The Structure-Mapping Engine: Algorithm and Examples'. *Artificial Intelligence*. 41: 1-63.
- Fauconnier, G. & Turner, M.** (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.
- Gentner, D.** (1983). 'Structure-Mapping: A Theoretical Framework for Analogy'. *Cognitive Science*. 7: 155-70.
- Harrell, D. F.** (2009). 'Toward a Theory of Phantasmal Media: An Imaginative Cognition and Computation-Based Approach to Digital Media'. *CTheory*.
- Humphrey, R.** (1954). *Stream of Consciousness in the Modern Novel*. Berkeley and Los Angeles: University of California Press.
- Johnson-Eilola, J.** (1997). *Nostalgic Angels: Rearticulating Hypertext Writing*. Norwood, NJ: Ablex Press.
- Mateas, M.** (2002). *Interactive Drama, Art, and Artificial Intelligence*. Carnegie Mellon University, Ph.D. dissertation.
- Meehan, J.** (1976). *The Metanovel: Writing Stories by Computer*. Yale University, Ph.D. dissertation.
- Woolf, V.** (1957 [1925]). 'Modern Fiction'. *The Common Reader: First Series*. V Woolf (ed.). London: Hogarth Press.
- Woolf, V.** (2002 [1925]). *Mrs. Dalloway*. Harcourt.
- Zhu, J. & Harrell, D. F.** (2010: forthcoming). 'Computational Narration of Inner Thought: Memory, Reverie Machine'. *Hyperrhiz: New Media Cultures*.
- Zhu, J. & Ontañón, S.** (2010). "Towards Analogy-Based Story Generation". *Proceedings of the First International Conference on Computational Creativity (ICCC X)*. Lisbon, Portugal, 2010.

Notes

1. Although this paper focuses on text-based narratives generated by computer algorithms, the core of the discussion can be extended to other computational narrative forms, such as video games.

Posters

An Approach to Ancient-to-modern and Cross-script Information Access for Traditional Mongolian Historical Collections

Batjargal, Biligsaikhan

biligsaikhan@gmail.com

Graduate School of Science and Engineering,
Ritsumeikan University, Japan

Khaltarkhuu, Garmaabazar

garmaabazar@gmail.com

Mongolia-Japan Center for Human Resources
Development, Mongolia

Kimura, Fuminori

fkimura@is.ritsumei.ac.jp

College of Information Science and Engineering,
Ritsumeikan University, Japan

Maeda, Akira

amaeda@is.ritsumei.ac.jp

College of Information Science and Engineering,
Ritsumeikan University, Japan

The main purpose of this research is to develop a system to keep over 800-year-old historical records written in traditional Mongolian script for future use, to digitize all existing records and to make those valuable data available for public viewing and screening. There are over 50,000 registered manuscripts and historical records written in traditional Mongolian script stored in the National Library of Mongolia. About 21,100 of them are handwritten documents. There are many more manuscripts and books stored in libraries of other countries such as China, Russia and Germany. Despite the importance of keeping 800-year-old historical materials in good condition, the Mongolian environment for material storage is not satisfactory to keep historical records for a long period of time (Tungalag, 2005). We believe that the most efficient and effective way to keep and protect old materials of historical importance while making them publicly available is to digitize historical records and create a digital library.

Mongolia introduced a new writing system (Cyrillic) in 1941. This was a radical change and alienated the traditional Mongolian script. The spelling of words and suffixes in traditional Mongolian script differs from spellings in modern Mongolian (Cyrillic). This is due to traditional Mongolian script preserving a more ancient language while modern

Mongolian reflects pronunciation differences in modern dialects. Spellings used in traditional Mongolian script reflect the Mongolian language spoken in the days of Genghis Khan but also contain elements of the ancient Mongolian language spoken before that era. Thus traditional Mongolian has a different grammar and a distinct dialect from modern Mongolian. At present, people use dictionaries with transcribing suffixes to overcome differences.

Recently ancient historical documents in traditional Mongolian script are being digitized and made publicly available, thanks to advances in innovative information technologies, and the popularity of the Internet. In Windows Vista, and later versions, especially in Windows 7, text-display support for traditional Mongolian script and the input locale is enabled. The Uniscribe driver was updated to support OpenType advanced typographic functionality of complex text layouts such as traditional Mongolian script. Traditional Mongolian script is written vertically from top to bottom in columns advancing from left to right.

However, retrieval of the required information in modern Mongolian from traditional Mongolian script documents is not a simple task, due to substantial changes in Mongolian language over time. We want to offer an information retrieval method that considers language difference over time. Our goal in this research is to develop a retrieval system where a user can access cross-period and cross-script databases with a query input in a modern language.

1. Related Research

Much research has been conducted in the last decade on Cross-language information retrieval (CLIR) – a technique to retrieve documents in one language through a query in another language. On the contrary, little research has been completed regarding information retrieval techniques for historical documents. Still less, almost none of the breakthroughs in research on information retrieval and information access have aimed at retrieving information in the native language from an ancient, cross-period and/or cross-script foreign language documents. Almost none of the CLIR systems has integrated either modern (Cyrillic) or traditional (ancient) Mongolian language due to its research backwardness.

Few approaches (Ernst-Gerlach et al., 2007; Garmaabazar et al., 2008; Kimura et al., 2009) that could be considered a cross-period information retrieval have been proposed. Little research has been completed regarding information retrieval techniques for historical documents. Ernst-Gerlach

et al. (2007) developed a retrieval method that considers the spelling differences and variations over time. They focused on modern and archaic German. Kimura et al. (2009) proposed a retrieval method that considers not only language differences over time, but also cultural and age differences in the same language. They focused on retrieval techniques for modern and archaic Japanese. Khaltarkhuu et al. (2008) proposed a retrieval technique that considers cross-period differences in dialect, script and writing systems of the same language. They focused on modern and traditional Mongolian.

2. System Architecture

We propose a simple model to cope with cross-period and cross-script queries. The proposed model is shown in Figure 1.

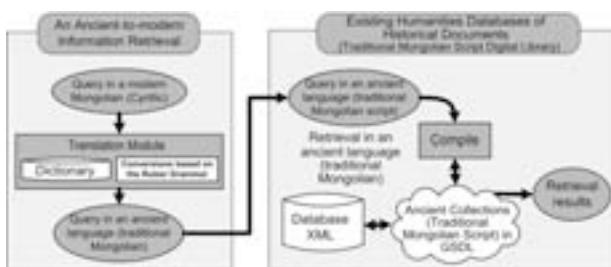


Figure 1

In the first stage, the query in a modern language will be translated into a query in an ancient form. Consequently, a query in the ancient language is submitted as a retrieval query for digital collections. The proposed model is expected to perform cross-period information retrieval and a user will be able to access ancient historical databases with a query input in a modern language. We will adopt a dictionary-based query translation approach.

3. Implementation of the Proposed System

Although it is not easy to develop such a retrieval system, we will utilize existing approaches (Garmaabazar et al., 2008; Kimura et al., 2009) to realize the proposed approach. A prototype called the Traditional Mongolian Script Digital Library¹ (TMSDL) (Garmaabazar et al., 2008), which could be considered a cross-period information retrieval system, has been developed. The retrieval method of the TMSDL considers cross-period differences in writing system of the ancient and modern Mongolian languages.

However, retrieval techniques of the TMSDL have not considered irregular words which have different meanings but are written and pronounced exactly the same in modern Mongolian and have different forms in traditional Mongolian. Also, word sense

disambiguation has not been considered. Thus, we enhanced the TMSDL by integrating a dictionary-based cross-period information retrieval approach. We utilized the developing online version of the Tsevel's concise Mongolian dictionary² (Tsevel, 1966) under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported license.

The Altan Tobchi (year 1604, 164pp) – A chronological book of ancient Mongolian Kings, Genghis Khan and the Mongol Empire (the largest contiguous empire in history) is shown in Figure 2, with modern Mongolian input interface in the new version of TMSDL (TMSDLv2). A database of such historical records in the TMSDLv2 with English or modern Mongolian query input will help someone conducting research in the history of the High Middle Ages, accessing materials written in an ancient language (traditional Mongolian) in order to understand 13th-14th century history of Asia.



Figure 2

4. Experiments

After the enhancement, we conducted a preliminary experiment to examine the difference between the TMSDL and the TMSDLv2 as well as to check the accuracy of translations from the modern language to the ancient one. We compared our retrieval results from the two versions with the "Qad-un ūndūsūn quriyangui altan tobči", (Textological Study) (Choimaa et al., 2002). This textological study contains the detailed analysis of traditional Mongolian words' frequency in the Altan Tobchi. All queries that were retrieved in the TMSDL were retrieved in the TMSDLv2. Therefore, we selected single word queries that were not retrieved in the TMSDL. In addition, selection criteria for the query input are:

- Pronounced or written differently in modern and traditional Mongolian; and

- With higher frequency in the Altan Tobchi.

In the experiment of retrieving traditional Mongolian documents via modern Mongolian (Cyrillic) utilizing a dictionary, we found that the TMSDLv2 translates and retrieves about 86% of selected input queries. However, about 64% of input queries have some variations that are less than or greater than the actual frequency because of the possible errors of translation and text digitization, or limitations of the retrieval function. We are working to distinguish the causes. Our translation module failed to translate 14% of input queries. Improvements on retrieval results are illustrated in Figure 3. Sample retrieval results are shown in Table 1. Retrieved results in the TMSDLv2 with highlights are shown in Figure 4.



Figure 3

A word in Modern Mongolian (Cyrillic)	Pronunciation		A word in Traditional Mongolian	Meaning, English translation	Retrieved results		Actual frequency of the word in the document	Retrieval rate
	Modern	Accent			Title TMSDL	Title TMSDLv2		
ийн	ийн	ийн	ийн	man, person, house	0	155	155	Fully retrieved
цар	цар	цар	цар	king	0	172	172	Fully retrieved
жар	жар	жар	жар	all, whole	0	0	0	Not retrieved
жад	жад	жад	жад	good, well, fine, nice, pretty	0	8	7	Not retrieved
чиг	чиг	чиг	чиг	person, human, creature, family etc., longitude, etc.	0	11	17	Greater than actual frequency
чаг	чаг	чаг	чаг	year	0	50	47	Greater than actual frequency
жин	жин	жин	жин	body	0	144	140	Less than actual frequency
жинэ	жинэ	жинэ	жинэ	body, second, second	0	34	43	Less than actual frequency

Table 1

Figure 4

5. Conclusion

In this paper, we proposed a model that utilizes cross-period and cross-script digital collections, which can be used to access old documents written in an ancient language using a query in a modern language. The proposed system is suitable for full text searches on databases containing cross-period and cross-script documents. Such research would involve extensive research in an ancient language that users and humanities researchers may or may not understand. It could apply to humanities researchers who are conducting research on ancient culture and looking for relevant historical materials written in that ancient language. The proposed model will enable users and humanities researchers to search for such materials easily in a modern language.

However, in our experiment, the TMSDLv2 translates and retrieves about 86% of input queries; only 22% is retrieved without any error. The other 64% has some differences on the number of retrieved query terms. For future development, improvements on translation and retrieval techniques need to be considered to increase the retrieval precision.

For future research, enhancements such as the retrieval of information from two distinct languages and retrieval via single query input from multiple humanities databases in multiple languages need to be developed. Our future work includes evaluating retrieval effectiveness by conducting extensive experiments that consider language differences over time. Performance of this approach will be compared with other approaches.

References

- Choimaa, Sh. and Shagdarsuren, Ts. (2002). *Qad-un ūndūsūn quriyangui altan tobči*,

(*Textological Study*). Ulaanbaatar: Centre for Mongol Studies, NUM.

Ernst-Gerlach, A. and Fuhr, N. (2007). 'Retrieval in text collections with historic spelling using linguistic and spelling variants'. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL2007)*. Vancouver, Canada, June 2007, pp. 333-341.

Garmaabazar, Kh. and Maeda, A. (2008). 'Developing a Traditional Mongolian Script Digital Library'. *Proceedings of the 11th International Conference on Asia-Pacific Digital Libraries (ICADL2008)*. Bali, Indonesia, December 2008, pp. 41-50.

Kimura, F. and Maeda, A. (2009). 'An Approach to Information Access and Knowledge Discovery from Historical Documents'. *Conference Abstracts of the Digital Humanities 2009 (DH09)*. College Park, MD, June 2009, pp. 359-361.

Tsevel, Y. (1966). *Mongol helnii tovch tailbar toli*. Ulaanbaatar (Mongolian).

Tungalag, D. (2005). *Mongol ulsiin undesnii nomiin san dahi Mongoliin tuuhiin gar bichmeliin nomzuin sudalgaa*. Ulaanbaatar (Mongolian). V. 1.

Notes

1. <http://www.dl.is.ritsumei.ac.jp/tmsdl>
2. <http://toli.query.mn/>

A Digital Archive of Buddhist Temple Gazetteers

Bingenheimer, Marcus

m.bingenheimer@gmail.com
Dharma Drum Buddhist College

Hung, Jen-jou

Dharma Drum Buddhist College

Temple gazetteers are a subset of the Chinese genre of gazetteers (*difang zhi* 地方志). Chinese gazetteers, or local histories, are composite texts containing descriptions, bibliographies, poems and other material pertaining to the history of a location or region. The temple gazetteers produced since the 16th century are important for the study of Chinese Buddhist history. They are especially relevant for the last three hundred years, but contain much older material on the history of Buddhist sites.

The archive is currently being constructed at the Dharma Drum Buddhist College, Taiwan, and for the first time opens up a large amount of this material for the study of Buddhism, which so far has been available only with great difficulty. The project website is at: <http://buddhistinformatics.ddbc.edu.tw/fosizhi/>.

The poster briefly introduces the content and then focusses on the technical realization of the "Digital Archive of Buddhist Temple Gazetteers".

Though mainly conceived as an online image database, the project includes 12 large gazetteers as full-text. These are marked up in TEI for names and dates and connected to the Buddhist Authority Database Project (<http://authority.ddbc.edu.tw>), to serve information to the interface. The project is an example for the growing trend to present full text next to digital facsimiles and the poster will show one way this can be done.

As of autumn 2009, the image database consists of more than 100,000 archival-quality images in TIFF format. The images were scanned in 8-bit greyscale at a resolution of 400dpi. From the digital master we produce watermarked JPEG files for use in the interface, and these are made freely available under a Creative Commons license. The quality of the JPEG files is sufficient to read and research the material.

The main limitation on the facsimile value of these images is that they are scans of copies of the original prints or manuscripts and therefore cannot achieve the same verisimilitude as facsimiles taken directly

from the originals. This deficit is mitigated by the fact that the material itself is unproblematic. Mostly text, it consists of black and white woodblock prints or brush writing. Due to the large character-size and the high image resolution the actual facsimile quality is high in the "Excellent Readability" range of the Quality Index (QI) benchmark for printed text (Kenney & Chapman 1995). Eventually, 237 gazetteers will be digitized and made available in the image database.

The gazetteer project includes various kinds of meta-data. We use MIX (Meta-data for still Images in XML) to record technical information about the image files, and meta-data for the TEI full-text files is kept in the teiHeader section of each file. While the MIX data is mainly generated automatically, the TEI meta-data is created as part of the mark-up process.

There are two other important datasets. First is the bibliographic data, which contains important information about the location of the temples, the relationship between the two printed collections, the edition history and chapter order of individual gazetteers, and additional bibliographic information on gazetteers collected from other sources. All this is kept in TEI files to allow for seamless integration with the other textual data.

Second, we have collected semantic data pertaining to each image file, i.e. each page. This data includes the image filename and the page number from the printed edition, which is required for the interface, especially for the majority of gazetteers which have not been digitized as full-text. This semantic image meta-data also records the first three characters on each page and, importantly, all title headings that appear on the page. This results in a database of all title headings, which in turn allows for shallow searches across the whole archive.

Although for production purposes it was useful to create distinct meta-data sources for the project, for archival purposes we integrate all of these different resources in METS wrappers.

The interface is based on the ExtJS JavaScript library (<http://extjs.com/>). ExtJS was chosen because it is one of the more advanced JS libraries available. It provides many vital interface functions 'out-of-the-box' and is used in for a number of other projects at the Library and Information Center. The full-texts are kept in a native XML database called eXist (<http://www.exist-db.org/>), which stores and retrieves the TEI source files.

The interface is geared to enable convenient reading of the gazetteers online. The design is based on the assumption that digital archives will increasingly want to present electronic text in conjunction with and alongside electronic facsimiles of the original

source. Among the challenges we solved here was how to have images and text move in unison. Another problem we faced was how to give the user an idea about where she was in the structure of the text, and it was decided to offer a detailed navigation panel, which contains detailed titles of every section to address this need. Finally, for better and more convenient accessibility of the image files we have included a magnifier to aid reading. A sophisticated search function lets users choose to search for strings in the fulltext, metadata or both, over the whole archive or one single gazetteer.

References

- Bingenheimer, Marcus** (forthcoming). 'Bibliographical notes on Buddhist temple gazetteers and some remarks on their use for the study of Chinese Buddhist history'. *Oslo Studies in Language*.
- Brook, Timothy** (2002). *Geographical Sources of Ming-Qing History*. Michigan monographs in Chinese Studies. Ann Arbor: Univ. of Michigan, Center for Chinese Studies V. 58.
- Eberhard, Wolfram** (1964). *Temple Building Activities in Medieval and Modern China*. Monumenta Serica. V. 23, pp. 264-318.
- Zhongguo Fosi Shizhi Huikan** (1980-1985). 中國佛寺史志彙刊. Compiled by Du Jiexiang 杜潔祥 (ed.). Taipei 110 vols.
- Zhongguo fosizhi congkan** (2006). 中国佛寺志叢刊. Compiled by Zhang Zhi 张智 (ed.). Hangzhou: Guangling shushe 广陵书社. 130 vols.

Preparing the Dariah e-Infrastructure

Blanke, Tobias

tobias.blanke@kcl.ac.uk

King's College London

Haswell, Eric Andrew

eah@hum.ku.dk

Nordisk Forskningsinstitut, University of Copenhagen,

With this poster, we would like to lay out our vision for Dariah and first steps towards its realization. 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 conceptualise and afterwards build a virtual bridge between different humanities and arts resources across Europe. Dariah is currently in its preparatory phase, which will design the infrastructure and build a sound business and governmental model. From 2011, Dariah will begin its construction phase.

Dariah starts off with the observation that just like astronomers require a virtual observatory to study the stars and other distant objects in the galaxy, 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. Its aim is to bring together various national infrastructures, such as the UK's arts and humanities e-Science initiative projects (<http://www.ahessc.ac.uk/>) and the German e-Humanities infrastructure TextGrid (<http://www.textgrid.de>). Dariah has also helped to found the Coalition of Humanities and Arts Infrastructures and Networks (CHAIN), an international group of arts and humanities infrastructure initiatives.

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 decentralised as possible, empowering individual contributors (e.g. individual researchers; national centers; specialised, thematic centers) to work with and within the Dariah community and shape its features as to their needs. Each contribution of each contributor builds Dariah, linked together in Dariah's architecture of participation. At the same time, however, collaboration across the borders of individual centers requires the usage of common technologies e.g. for authentication or federation of archive contents.

Dariah is about the (*re-*)use of digital research resources by anybody and anywhere. By providing data for anybody's use, it is an e-Research environment. With regard to standards to foster interoperability of content and compatibility of tools, Dariah will not prescribe but encourage. Researchers do not *have to* support interoperability and openness, but they *may want* to support and benefit from opportunities such as collaboration and re-usability facilitated by interoperability and openness. Dariah provides community-driven recommendations and fosters interoperability and collaboration through incentives. This approach means less central control over what Dariah contains and provides.

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

We imagine Dariah therefore to be not one large infrastructure but more a means of linking up people, services and data for research in arts and humanities.

Most likely, Dariah will not be one technical solution but many, according to community needs and willingness to collaborate. And it is in this context that engaging with the active and vibrant community of international digital humanists is high on Dariah's list of priorities. Dariah is engaged in important, in-depth work in modelling research needs and behaviours, the results of which will inform the further development of Dariah.

We think that the definition in the DuraSpace midterm report on what is a repository also fits Dariah: "trusted intermediary that makes content (...) usable with a menu of added-value services". Of course Dariah will not be one large repository, but otherwise the idea of a trusted intermediary fits well.

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.

Cultures of Knowledge: An Intellectual Geography of the Seventeenth-Century Republic Letters

Brown, James

james.brown@history.ox.ac.uk
University of Oxford, UK

Hotson, Howard

University of Oxford, UK

Jefferies, Neil

University of Oxford, UK

This combined poster and software demonstration will introduce 'Cultures of Knowledge: An Intellectual Geography of the Seventeenth-Century Republic Letters', launched in January 2009, and based in the Humanities Division of the University of Oxford with funding from the Andrew W. Mellon Foundation. Comprising a diverse group of academics and technical experts from the Faculties of History, English, Theology, and Bodleian Libraries, as well as from partner institutions in Britain and east-central Europe, the Project is seeking to reconstruct the correspondence networks that were central to the revolutionary intellectual developments of the seventeenth century. One group is working to catalogue, edit, and preserve the rich archives of scientific correspondence deposited in the Bodleian Library. Another is working with colleagues in Sheffield, Prague, Cracow, and Budapest to enhance and link letter collections elsewhere. Finally, a third group, based in the Systems and e-Research Service (SERS) of Bodleian Libraries, is developing a digital system capable of organising metadata on these materials into an online union catalogue of intellectual correspondence. It is this central feature of our collaboration that we wish to share at DH2010.

While meticulously edited and annotated hard copy editions of finite corpora remain indispensable, the intellectual and methodological imperatives for looking beyond traditional modes of publication are overwhelming in the context of the Republic of Letters. No correspondence is an island unto itself. On the contrary, every single letter links at least two different correspondence networks: that of its sender, and that of its recipient. Many seventeenth-century letters passed through multiple hands, either because they were written by or addressed to more than one person, because they passed through intermediaries, or because they were (re)circulated through broader circles of friends and associates

after receipt. Moreover, the provision of increasingly frequent, fast, and inexpensive postal services – what has been described as the early modern European ‘media revolution’ – meant that these systems of connections extended over enormous distances. Traditional editions of letters to and from single eminent individuals fail to capture the multilateral, spatially-dispersed character of early modern epistolary cultures, and additional tools are needed in mapping the broader networks which surrounded these canonical figures (networks which are becoming central to research in the history of science and related fields).

To this end, capitalizing on the unprecedented opportunities created by ‘digital revolution’ of recent decades, we are proposing to create the nucleus of a web-mounted union catalogue of seventeenth-century correspondence which, for the first time, will provide scholars with a means of navigating the vast, uncharted sea of correspondence that surrounded the comparatively well-surveyed islands of seventeenth-century intellectuals. In its first phase, it will combine three large datasets generated by the Project or already hosted by the Bodleian Library, namely: a digitized version of the existing Bodleian catalogue of seventeenth-century manuscript correspondence (c.26,000 letters); digital calendars of six discrete correspondences generated by our individual research projects (cataloguing the letters of John Aubrey, Jan Amos Comenius, Samuel Hartlib, Edward Lhwyd, Martin Lister, and John Wallis [c.10,000 letters]); and metadata from the seventeenth-century correspondence already included in the Electronic Enlightenment (c.7,000 letters).¹ This will result in an integrated database of over 40,000 letters which will provide both a platform for our own future efforts in the field and, it is hoped, a critical mass of material necessary to attract further contributions from elsewhere.

The poster will outline the technological as well as the conceptual underpinnings of this enterprise. We will describe the system architecture of the union catalogue, the Digital Asset Management System (DAMS), an innovative platform developed to support digital library projects within the University of Oxford. The DAMS, based on Fedora software, provides a robust and flexible mechanism for the storage of digital objects within an RDF/semantic web framework. Predicated upon adaptability, scalability, interoperability, and long-term preservation, the system is particularly suitable for our purposes. It allows content to be openly accessible to, and available for reuse by, a geographically dispersed scholarly network; will enable us to share data with electronic repositories elsewhere (and vice versa); and will ensure that

the union catalogue will remain accessible to international research long beyond the lifecycle of our original grant. We will also describe the metadata standards by which individual correspondences will be styled as Fedora objects within the system, the RDF ontologies by which they will be related, and the software for web-based record editing, as well as data collection and import, which has been developed especially for the Project. During the poster session itself we will provide a live demonstration of a pilot implementation of the union catalogue. This will showcase its key features (full-text faceted search engine; predefined browsable views; links to select transcriptions and digital images; look-and-feel), as well as the interface for online editing that will allow individual records to be amended, and new records added, by a widely distributed community of academic contributors.

Key Contacts

Professor Howard Hotson (Project Director):
howard.hotson@st-annes.ox.ac.uk

Neil Jefferies (R&D Manager, SERS):
neil.jefferies@sers.ox.ac.uk

Ben O’Steen (DAMS Architect, SERS):
benjamin.osteens@sers.ox.ac.uk

Sue Burgess (Project Systems Developer, SERS):
sushila.burgess@sers.ox.ac.uk

Dr James Brown (Project Coordinator):
james.brown@history.ox.ac.uk

Further Information

<http://www.culturesofknowledge.org>

More bibliography at <http://www.history.ox.ac.uk/cofk/bibliographies>

References

Awre, Chris, Swan, Alma. 'Linking Repositories: Scoping the Development of Cross-Institutional User-Oriented Services'. *OCLC Systems & Services: International Digital Library Perspectives*. **23** (2007): 372–81.

Berkvens-Stevelinck, Christiane, Bots, Hans, Haeseler, Jens (eds.) (2005). *Les grands intermédiaires culturels de la République des Lettres: Etudes de réseaux de correspondances du XVIe au XVIIIe siècles*. Paris.

Muchembled, David, Bethencourt, Francisco, Egmond, Florike (eds.) (2007). 'Cultural Exchanges in Early Modern Europe'. *Correspondence and Cultural Exchange in Europe, 1400–1700*. Cambridge. 3 vols.

Bots, Hans, Waquet, Françoise (eds.) (1994). *Commercium litterarium, 1600–1750. La communication dans la République des lettres/Forms of Communication in the Republic of Letters.* Amsterdam.

Davies, John, Fensel, Dieter, van Harmelen, Frank (2003). *Towards the Semantic Web: Ontology-Driven Knowledge Management.* Hoboken, NJ.

Duraspace. <http://duraspace.org/index.php>.

Grafton, Anthony (2009). *Worlds Made by Words: Scholarship and Community in the Modern West.* Cambridge, MA.

Jaumann, Herbert (ed.) (2001). *Die europäische Gelehrtenrepublik im Zeitalter des Konfessionalismus.* Wiesbaden.

Nichols, Stephen. 'Time to Change our Thinking: Dismantling the Silo Model of Digital Scholarship'. *Ariadne.* <http://www.ariadne.ac.uk/issue58/Nichols/>.

RDF. <http://www.w3.org/RDF/>.

Notes

1. See <http://www.e-enlightenment.com/>

Supporting User Search for Discovering Collections of Interest

Buchanan, George

g.r.buchanan@gmail.com

School of Informatics, City University, London

Dodd, Helen

cshelen@swansea.ac.uk

Future Interaction Technology Group, Swansea University, Swansea

Humanities researchers often draw their information from a diverse set of sources, which are unlikely to be found in one digital library or other collection. Whilst a considerable volume of technical work has been undertaken to unify a number of digital libraries into one whole - "federated" or "distributed" digital libraries - the real-world adoption of this technique is riddled with conceptual problems and organisational practicalities. In consequence, there is little likelihood of there arising one "super-gateway" to which an active researcher in the humanities can turn for the whole of their information seeking. Indeed, even in the sciences, this idealised situation is relatively rare. As a result, a humanities researcher will often need to identify a number of different collections that serve their regular information needs well [Buchanan *et al.* 2005]. Each of these collections would be likely to provide useful literature relevant to their field of study. However, constructing this list is problematic. At present a user can only generate such a list of 'good' collections from months and years of incidental discovery and recommendation.

In the past, this need has been addressed in a number of different ways. The Humbul humanities hub - later part of Intute - was started as a human-maintained list of online resources for the humanities and arts. It relied upon hand-crafted entries created by researchers (often postgraduate students) that described each collection and suggested its potential uses. One positive advantage of this method is that it develops a single list of many collections, and is open-ended. However, there are problems that emerge when a user tries to find collections within the hub. This approach necessarily requires the creator of an entry to double-guess the likely tasks of another user and a substantial cumulative effort over many years. It is resource expensive in terms of creation, difficult to maintain, and an entry is very unlikely to be able to cover all likely uses in exactly the terms that another researcher may use. Whilst a positive benefit is that the system provides a 'human readable'

overview of each site, if a user provides a search to the system, it will only use the relatively small amount of material entered by the researcher who entered the site onto the system. Compared to the information available within a single library, a brief descriptive entry is prone to have insufficient information, and may overlook the searcher's interest altogether, or provide a disproportionate representation - greater or lesser - of the volume of material relevant to their work.

In contrast, highly technical approaches have been undertaken to create a central, canonical resource through "metasearch" techniques: where a central service endeavours to offer a synthetic unification of a number of DL systems [Thomas and Hawking 2009]. In these methods, the user provides a sample query that is then automatically sent to every individual digital library. The metasearch engine combines the results from the separate libraries into one whole, and returns the unified set of results to the user.

One limitation to the metasearch approach is that the list of libraries or search engines supported is usually fixed. Metasearch on the web often relies on "reverse-engineering" the HTML output from each single search facility that the metasearch system uses. This requires extensive maintenance work, and lists of available collections are thus often fixed.

Another concern with this approach is that from a conceptual view, it is vulnerable to poor understanding or even utter ignorance of the operation of each constituent DL. Different DLs may interpret the same search in radically different ways. Another issue is that as these services normally send a search to each of their constituent DLs for each search given by a user, and this means a substantial overload of search activity for each constituent DL. Practically, this is clearly ineffective and costly in terms of computation and, ultimately, hardware resources.

One method proposed to minimise such waste is to provide a "database selection" algorithm that pre-selects only the better DLs to search, and these are then queried by the metasearch system automatically [Thomas and Hawking 2009, French *et al.* 1999]. The metasearch system then combines the result sets from each chosen DL and provides a single ranked list of matching documents. However, the same problems with general metasearch return: how results are combined into one remains an issue, and users are known to be poor at reconstructing the best sources (i.e. the sites with the largest volume of relevant material) from such lists.

In our research, we do not attempt to circumvent the conceptual problems of unifying search result lists from different libraries with varying matching algorithms and heterogenous vocabularies. Rather,

we aim to embrace the diversity of information sources, and suggest to a user the libraries that are more likely to contain good-quality information for a given query. In this regard, our approach is similar to meta-search. However, there are key differences in our method: first, we provide the user with a list of likely "best" DLs, rather than individual documents; second, we do not attempt any merging of result lists or other conceptually fraught manipulation of retrieved material; thirdly, use DL technologies such as the OAI protocol for metadata harvesting (OAI-PMH) to provide an open-ended list of target collections [Van de Sompel *et al.* 2004], in contrast with the fixed list limitation that is commonplace in metasearch; fourthly, we are reconsidering the best matching algorithms from a user-centred perspective, rather than from currently commonplace information-retrieval based metrics that may poorly match a user's requirements when their aim is to discover rich, large-scale sources of information on a particular topic.

Our current research has probed this fourth and final issue. What we have uncovered is that many of the current database selection algorithms privilege two sorts of collections: first, large collections; second, collections in which a sought-for term is rare. There are good reasons why this is appropriate, from a technical information-retrieval point of view. For example, a term that provides strong discrimination within a given collection is likely to produce a clear, consistent list of matching documents. However, from the perspective of a humanities researcher investigating a new topic, and who is seeking libraries with good coverage of the topic, these extant IR measures seem a poor match against their requirements. A digital library that contains a high proportion of documents with the sought-for topic is arguably very likely to be of long-term value to them. However, this pattern is exactly the opposite to that desired by the existing database selection algorithms, where a strongly distinctive term - i.e. one that matches only a few documents - is ideal. Similarly, absolute size is not necessarily a criterion. Often, a small but specialised high-quality collection is of critical value in orienting the humanist in a new domain. Furthermore, our own previous research suggests that humanities researchers do not trust computer-based models of relevance: they prefer computer systems to err slightly on the side of deferring such decisions to the user.

In assessing the current state-of-the-art, we have developed an experimental apparatus that permits the testing of several algorithms in parallel. Conceptual, idealised scenarios can be test-run, and then the same tests applied to sets of collections gathered by the researcher to retest the same scenario on real data. This permits us to assess any algorithm

against nominal and real data, and against a number of alternatives. This test environment has already demonstrated that many current database selection algorithms perform very poorly against the ideal criteria for the task that we seek to support, and that even the best are far from optimal. The bias towards large collections outlined above has been reiterated in practice, and we have also uncovered the problem that the numerical ratings produced by good algorithms follow unhelpful patterns.

One frequent problem is that subtly different scores for different collections against a particular search can be produced from profoundly different underlying coverage of search terms in the collections. This underlying problem is manifested in different ways. Common oddities include relatively high scores being given when only one term is matched (albeit many times) and ‘normalisation’ of scores meaning that very different matches between collection and query result in marginally different final scores. This second problem means that common methods for deciding on a list of ‘best’ matches do not work, and it is difficult to decide the criteria to use to interpret a score into a final recommendation.

A considerable body of further work is required. Whilst we are confident, from the current literature, that our current results are, for idealised scenarios, closer to what is required, the problem that we seek to answer is as yet poorly understood. We need to investigate further not only the technology, but also the human context in which it will operate, and through this develop a more sophisticated and accurate model of what humanities researchers would ideally require as the output of our system.

in information Retrieval (SIGIR '99). Berkeley, California, United States, August 15-19, 1999. New York, NY: ACM, pp. 238-245.

Van de Sompel, Herbert, Nelson, Michael L., Lagoze, Carl, Warner, Simeon (2004). 'Resource Harvesting within the OAI-PMH Framework'. *D-Lib Magazine*. **12**. http://www.dlib.org/dlib/december04/van_desompel/12vandesompel.html.

References

Buchanan, George, Cunningham, Sally Jo, Blandford, Ann, Rimmer, Jon, Warwick, Claire (2005). 'Information Seeking by Humanities Scholars'. *9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*. Vienna (Austria), September 18-23, 2005, pp. 218-229.

Thomas, Paul, Hawking, David (2009). 'Server selection methods in personal metasearch: a comparative empirical study'. *Information Retrieval*. **5**: 581-604.

French, James C., Powell, Alison L., Callan, Jamie, Viles, Charles L., Emmitt, Travis, Prey, Kevin J., Mou, You Y. (1999). 'Comparing the performance of database selection algorithms'. *Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development*

An Inter-Disciplinary Approach to Web Programming: A Collaboration Between the University Archives and the Department of Computer Science

Janet Marie Bunde

bunde@nyu.edu
New York University, USA

Deena Engel

deena@cs.nyu.edu
New York University, USA

Computing in the Humanities, an undergraduate course at New York University, represents a unique collaboration between the Computer Science Department and the University Archives. The final assignment required students to select, digitize, and contextualize materials from the Archives' collections in an interactive website. The design and implementation of the course incorporates four current and important trends in both disciplines. First, the professor and archivist worked closely together before and during the course, integrating the archival research component into the core mission of the course. Second, the students' projects provided both increased subject access and dynamic Web content to the repository and meaningful work to the students. Third, this course produced students who "bridge" the needs of humanists with the capabilities of technology. Fourth, this course illustrated the growing importance of web programming in undergraduate computer science education.

1. Trend 1: Archives and archivists in the undergraduate classroom

Archiving journals and conference presentations have increasingly focused on the experiences of archivists at educational institutions who seek to integrate archival resources into the undergraduate curriculum. Notably absent from this literature, however, is any mention of collaboration between archivists and professors in computer science or related departments.

Before the course began, the professor, undergraduate librarian, and University archivists

discussed the structure of the course. The librarian and assistant University archivist co-lectured for one class session. The students who elected to work in the University Archives for their final projects were also required to meet with the assistant archivist before embarking on their research. These meetings ensured that students would have sufficient resources at the Archives to conduct their primary (visual) and secondary (textual) research. Students from the class were also required to digitize photographs, documents, and audio materials as part of the assignment.

At the end of the semester many students who had used the University Archives for their project turned in their assignments for publication on the University Archives website. In doing so, the difference between a class project and a project published on the Internet under the authority of the University Archives became clear.

2. Trend 2: Projects increase access to archival materials and provide meaningful work for students

The projects produced by students in this course reflected the students' interests in the University Archives' collections—interests that do not necessarily coincide with the digitization and processing priorities of Archives staff. Archives staff can use this information to determine what researchers might look for in our collections and use this information to influence future digitization and processing priorities. Additionally, by digitizing materials from across collections centered on a particular theme, the students actually increased access for other users.

By allowing students to curate their own mini-collections, the repository opened up new opportunities to interpret these materials. This interpretation represents an instance of what Tom Nesmith describes as the "new description... a much more thorough contextual mapping of pathways through the masses of records [that] add[s] evidential value to the records" (Nesmith, 2007). By juxtaposing student projects with finding aids prepared by the Archives, a viewer of the University Archives website can compare the students' interpretations and the context of the records that were digitized. This contextualization is essential to understand these projects and the records they feature rather than provide additional "background noise of the World Wide Web" (Eamon, 2006). Future collaborations might include students working with the Archives-generated XML itself to discover new methods of embedding information,

or modifying the schema or stylesheets to include different information.

3. Trend 3: Nexus between computer science and archives

The course introduced new users to the repository and creates new areas of collaboration between archivists and faculty. Students enrolled in the course learned rudimentary archival research skills and, most importantly, were introduced to the same issues archivists wrestle with on a daily basis when providing materials to users.

Many articles written in the past decade mention how online digital collections and delivery of finding aids online have become rote rather than exceptional for most repositories. Yet at the same time there have been few overtures by archivists to bridge this digital divide at the undergraduate or even at the graduate level. In her 2005 survey of archival job postings, Michelle Riggs found that employers increasingly require knowledge of EAD and the markup languages HTML, XML, and SGML. She urges archival education programs and library science programs to offer instruction that matches employer demand (Riggs, 2005).

By introducing students with programming skills to research in an archive, this course also sought to create students who understand and can combine the capabilities of technology with the needs of a researcher. Corinne Jörgensen laments that the professionals who create access systems for digital resources, and those who access those resources, do not speak to each other or reference each other's research (Jörgensen, 1999).

4. Trend 4: Emerging importance of web programming in undergraduate computer science education

The Computer Science Department undergraduate program at New York University offers both a Computer Science major and minor as well as a minor in Web Programming and Applications. The department encourages faculty to create courses in web programming to meet the needs of students in both the CS major and minor programs.

Web programming has been increasing in importance in university Computer Science Education. "Despite its reputation in some circles, web programming is conceptually deep; it gives a simple way to learn event-driven programming, to become conversant in many languages, learn the client-server paradigm, interact with databases,

and more" (Stepp, 2009). Students with only one semester of studies in implementing websites and one semester of a high-level programming language such as Java or Python can build complex and interactive websites. We can thus reach students across a variety of disciplines, as this project offers a rich opportunity for inter-disciplinary studies. Both early CS majors and CS minors have an opportunity to focus on the content of the sites as well as the technology and programming required to build them.

Programming and technology requirements for this project included:

1. XML and XSLT (using text editors without a WYSIWYG interface) for the collection catalogues
2. PHP and JavaScript as well as advanced XHTML and CSS for the user interface
3. Original podcasts and work with appropriate multi-media objects related to the collections

At the University of Houston, Clear Lake, faculty used the following goals to design programming assignments for web programming coursework (Yue, 2004):

- Realistic: the assignments should be similar to useful real-world projects.
- Complete and ready: the products of the assignments should be Web applications that can be deployed with no or little modification.
- Technically important: the assignments should use important concepts and technologies.
- Illustrative and interesting: the assignments should be intellectually appealing and interesting.

We believe that the project that we developed together for this course meets these criteria with the added benefit of the intellectual and research challenge posed to the students using primary source materials as the foundation of their work. This allowed us to expand on the criteria as follows for our projects:

- The assignments should provide a basis for substantive historical research.
- The assignments should provide a basis for interdisciplinary discussion and research for dual-major students and students with majors outside of computer science.
- The assignments should provide real world experience with skill sets that students in the Humanities fields would use in future research and graduate studies.
- The assignment should provide the student with an opportunity to experience the complete life-cycle of a programming project: from inception,

research and design through implementation and publication.

In conclusion, we believe that the current trends in archival research and computer science undergraduate education have converged in a way that provides rich opportunities for our students as well as for the university. Students were engaged in and enthusiastic about their projects, which represented real-world applications of the concepts that comprised the course's goals. Several students cited the course as their inspiration to attend graduate school in library and information science. We believe that the need for graduates with this combined set of skills will continue to grow.

References

- Eamon, Michael** (2006). 'A 'Genuine Relationship with the Actual': New Perspectives on Primary Sources, History, and the Internet in the Classroom'. *The History Teacher*. **39**, 3: 1-32.
- Jørgensen, Corinne** (1999). 'Access to Pictorial Material: A Review of Current Research and Future Prospects'. *Computers and the Humanities*. **33**: 293-318.
- Nesmith, Tom** (2007). 'What is an archival education?'. *Journal of the Society of Archivists*. **28**, 1: 1-17.
- Riggs, Michelle** (2005). 'The Correlation of Archival Education and Job Requirements Since the Advent Encoded Archival Description'. *Journal of Archival Organization*. **3**, 1: 61-79.
- Stepp, Marty, et al.** (2009). 'A 'CS 1.5' Introduction to Web Programming'. *Proceedings of SIGSCE*. Pp. 121-125.
- Yue, Kwok-Bun, Wei Ding** (2004). 'Design and Evolution of an Undergraduate Course on Web Application Development'. *Proceedings of ITICSE*. Pp. 22-26.

Citation Rhetoric Examined

Dobson, Teresa M.

teresa.dobson@ubc.ca

University of British Columbia, Canada

Eberle-Sinatra, Michael

michael.eberle.sinatra@umontreal.ca

Université de Montréal, Canada

Ruecker, Stan

sruecker@ualberta.ca

University of Alberta, Canada

Lucky, Shannon

lucky.shannon14@gmail.com

University of Alberta, Canada

INKE Research Group

inke.project@gmail.com

INKE Project

In his influential monograph *The Rhetoric of Citation Systems*, Connors (1999) elaborates on the principle that scholars working with different forms of citation find themselves thinking differently, since the citation format has natural consequences in the way it interacts with the material in the practice of the writer. For example, the popular MLA and APA formats differ radically in the way they handle footnotes. MLA allows writers to include both substantive and citation footnotes, and gives them the choice to include citations at the foot of the page, at the back of the book, or inline. Many journals employing APA, on the other hand, discourage use of substantive footnotes and require that citations be inline. The content of in-text parenthetical citations is also different: MLA requires writers to include a page number for citations, while APA allows writers to refer broadly to a source by author name and year. Connor argues that the APA's emphasis on the year encourages both the writer and the reader to be conscious of how recent the source material is, and that a prejudice tends to emerge against older publications, which helps to strengthen the supercessionist form of thinking across the disciplines where the APA format is popular. In addition, the lack of substantive footnoting in the APA tends to discourage digressions into related but essential content by both writers and readers.

In this project, we examine how citation style may modify the reading experience through a preliminary study with graduate students in the humanities and social sciences. In the first half of the study, we asked five graduate students to read an article that contains a number of substantive footnotes

(Booth's "Witchcraft, flight and the early modern English stage"), and had them prepare an alternative version in which they were required to incorporate all footnotes into the text. These two versions correspond roughly to the practice in MLA and APA citation. We note, for instance, that MLA does not require substantive footnoting, and APA does allow for substantive footnotes; however, in practice footnotes are more widely accepted by journals employing MLA style, while APA journals often discourage or disallow footnotes. The goal of the exercise was to look at the differences in handling the two conditions and the effects of those differences on production and reception of academic content. We recorded the details of this process employing Morae usability software and interviewed participants post-task about their process. The study participants found this to be a challenging and at times frustrating exercise. Many remarked that the information in the discursive footnotes was extraneous to the main thesis of the article and that there could be no satisfactory way of integrating that information into the text. Some participants included the footnoted material verbatim parenthetically in-text. Others omitted the footnotes altogether. Generally, participants felt that a paratextual space for discursive content (footnotes) is important, although in the case of this article the footnotes may have contained more extraneous material than would be desirable. Participants found converting in-text parenthetical references more straightforward. A prejudice against the inclusion of the year of publication in text in accordance with APA guidelines emerged: participants noted that the year of publication is less important to understanding the relevance of cited sources than other descriptive information such as title. The respondents did not believe either citation system was better equipped to help readers locate referenced sources, but they did indicate that discursive footnotes provide an important venue for valued parallel discussions and related but non-essential information, and that discouraging the use of discursive footnotes impoverishes academic writing.

Subsequently, we gave the same article to twenty-six undergraduate students: half of the participants read the original, MLA, version of the article; the other half read an APA version of the same article in which footnotes were integrated into the text. Post-reading, participants answered a comprehension question and nine recall questions. Citation style did not appear to affect either comprehension or recall. A significant trend, however, emerged: discursive information, whether it was located in the footnotes in the original MLA version or integrated into the text in the converted APA version, was far less likely to be recalled by participants. More than two thirds of participants did not answer questions based

on discursive information correctly, regardless of whether that information was contained in a footnote (as in the original MLA version) or integrated into the text (as in the prepared APA version). This finding appears to temper Connor's thesis about the effects of citation formats on reading.

This project is an initiative of a major collaborative research initiative in the digital humanities, Implementing New Knowledge Environments (INKE), that aims to foster understanding of the significance of digital and analog books and their role in humanities scholarship. It is also part of a larger study of citation rhetoric as exemplified in *Synergies: Canada's Social Sciences and Humanities Research Infrastructure*, a not-for-profit platform for the publication and dissemination of research results in the social sciences and humanities published in Canada. Results of this citation rhetoric research project will benefit analysis of citation statistics in large-scale web search interfaces such as *Synergies*. It will also contribute to further research on automated semantic searches in bibliographies and works cited, such as those covered in the first year of the INKE project (e.g. Ruecker et al. 2009).

References

- Booth, Roy** (2007). *Witchcraft, flight and the early modern English stage*. Early Modern Literary Studies.
- Connors, Robert J.** (1999). *The Rhetoric of Citation Systems*. NY: Lawrence Erlbaum Associates (Taylor & Francis Group).
- Ruecker, Stan, Rockwell, Geoffrey, Radzikowska, Milena, Sinclair, Stéfan, Vandendorpe, Christian, Siemens, Ray, Dobson, Teresa, Doll, Lindsay, Bieber, Mark, Eberle-Sinatra, Michael, NKE Group.** 'Drilling for Papers in INKE'. *INKE 2009: Research Foundations for Understanding Books and Reading in the Digital Age*. 23-24 October 2009.

Evidence of Intertextuality: Investigating Paul the Deacon's *Angustae Vitae*

Forstall, C. W.

forstall@buffalo.edu

Department of Classics, State University of New York at Buffalo

Jacobson, S.L.

slj25@buffalo.edu

Department of Classics, State University of New York at Buffalo

Scheirer, W. J

wjs3@vast.uccs.edu

Department of Computer Science, University of Colorado at Colorado Springs

The study of intertextuality, the shaping of a text's meaning by other texts, remains a laborious process for the literary critic. Kristeva (Kristeva, 1986) suggests that "Any text is constructed as a mosaic of quotations; any text is the absorption and transformation of another." The nature of these mosaics is widely varied, from direct quotations representing a simple and overt intertextuality, to more complex transformations that are intentionally or subconsciously absorbed into a text. The burden placed upon the literary critic to verify suspected instances of intertextuality is great. The critic must reference a large corpus of possible contributing works, and thus must often be familiar with more texts than was the author whom they are studying. Since, in many cases, the problem is one of pattern recognition, it is a good candidate for automated assistance by computers.

In this work, currently in progress, we propose the use of machine learning and related statistical methods to improve the process by which intertextuality is studied. Specifically, we are working with instances where an author has knowledge of particular texts, and reflects this in discrete passages within their own work. These intertexts may comprise fragmentary quotations, paraphrases, or even stylistic similarity. A passage may be reminiscent of a particular author, or of a particular literary group. We have defined three different classes of style markers to verify intertexts: phonetic, metric, and dictional. To evaluate the proposed style markers and classification methods, we have chosen an intriguing case study: Paul the Deacon's 8th century poem *Angustae Vitae* (Paul the Deacon, 1881), which we suggest has a strong connection to first-century Neoteric poetry.

Our earlier work in authorship and stylistic analysis (Forstall and Scheirer, 2009) has considered the importance of phonetic style markers, with the observation that sound plays a fundamental role in an author's style, particularly for poets. To capture sound information, we have developed a feature that we call a *functional n-gram*, whereby the power of the Zipfian distribution is realized by selecting the n-grams that occur most frequently as features, while preserving their relative probabilities as the actual feature element. By using more primitive, sound-oriented features, namely, character- and phoneme-level n-grams, we are able to build accurate classifiers with the functional n-gram approach. We have used two different classification algorithms with functional n-grams, yielding very promising results. The first method, a traditional SVM learning approach based on the work of Diederich et al. (Diederich et al., 2003), distinguished authors of Latin poetry with 98.75% accuracy. The second method, a PCA clustering approach (Holmes et al., 2001), showed distinct stylistic separation between the Homeric poems. In light of our previous work, we know that phonetics is an important tool for verifying intertexts, for the same reason it was important for poetics — repetitive sound distinguishes style.

Following this idea of repetitive sound further, we use meter as an additional style marker. For strict meters, it is straightforward to identify their type by analyzing the weights of the syllables in a line. In practice, the nuance of particular poets, or groups of poets, creates unique variations in meter, giving us a discriminating feature. By including meter as another dimension in the feature vector of the SVM learning for the functional n-grams described above, we enhance the discriminatory power of the resulting classifiers. It remains an open question in our work whether meter alone is powerful enough to achieve the same classification results for individual authors as the functional n-gram. Its utility for group classification is more apparent.

Pulling back from sound, we have also developed a style analysis for diction that is somewhat the opposite of the functional n-gram approach. Considering the Zipfian distribution once again, we turn to elements that occur with lower probabilities. The power of functional n-grams relies on the amount of information carried by the elements at the left side of the Zipfian distribution (assuming the x axis is organized from most frequent to least frequent). For this new style marker, we desire something that is the opposite of functional—features that occur infrequently, but not necessarily *hapax legomena*.

Thus, we fix a desired probability range for words that occur infrequently, and look for n-gram sequences

composed of only those words in a particular passage, ignoring all others:

$$(P_{low} < \text{Pr}(\text{word}_1) < P_{high}) \dots (P_{low} < \text{Pr}(\text{word}_2) < P_{high}) \\ \dots (P_{low} < \text{Pr}(\text{word}_n) < P_{high}) \quad (1)$$

where $n \geq 3$. The probability of the resulting n-gram is compared to pre-computed probabilities of the same n-gram (should it exist) for specific authors, or literary groups. This type of style marker is very well suited to our case study, where certain word sequences are common to a particular group (the Latin Neoterics), but uncommon or non-existent in the work of other groups.

With our style markers in hand, we turn to our case study. In *Angustae Vitae*, Paul the Deacon opposes poetic inspiration and production in the classical world with the writing of poetry in the Christian monastic context. Although he posits the classical and monastic worlds as opposites, the use of Catullan diction and models of poetic exchange recalls the paradigm of the Neoteric, proto-elegiac lover, his beloved, and his poetological concerns. This model is recontextualized to reflect monastic love and poetic exchange. The source of inspiration remains the same — love — but there is a new beloved and a new Muse. While he avoids saying he directly imitates his classical predecessors, Paul the Deacon's poetry is peppered with classical intertexts. Not all of these intertexts are purposeful allusions. However, it is clear that he was at least well-versed in the poetry of Horace, Virgil, and Ovid. Thus, it becomes our task to verify the portions of the poem believed to be inspired by Catullus.

Further study of the Catullan manuscript tradition and Paul the Deacon's life would be necessary to prove his knowledge of Catullus conclusively. This work will proceed from the *a priori* conclusion that Paul the Deacon had read Catullus. This conclusion is based on the abundance of intertexts and the crucial role they play in coloring Paul the Deacon's poetry. We can gain a sense of these intertexts we are examining by looking at a particular instance.

As the poem opens, the Muses, who have found the cloistered life not of their liking, have abandoned the narrator. More precisely, the Muses are fleeing the fellowship of the cloistered life, *angustae vitae fugiunt consortia Musae*. Thus, there is an opposition: the *consortia angustae vitae versus the Musae* — the fellowship of monastic life versus the classical relationship of the poet and Muses. This opposition continues in lines 2 — 3. The Muses do not wish to live in the fenced-in gardens of monasteries, but rather they desire to play in rosy meadows, *clastrorum septis nec habitare volunt, / per rosulenta magis cupiunt sed ludere prata*. Here, *septa*, the cultivated, enclosed garden of a monastery,

is contrasted with the *rosulenta prata*, a wild, open meadow.

These opening lines may reference Eclogue 1 of Vergil, but reading them with Catullus provides a richer understanding of the themes of the poem. Indeed, what the Muses desire in *Angustae Vitae* are cornerstones of Catullan diction. Lines 2 — 3 recall the opening lines of Catullus 2. The Muses desire to play in fields (*cupiunt sed ludere prata*) and to tend to (*colunt*) their delights (*delicias*), just as Catullus' girlfriend is accustomed to play (*ludere solet*) with her own pet/delight (*deliciae*). The classical Muses are compared with the poet's beloved. These are the Muses of elegiac love. *Ludere*, furthermore, is a by-word in Catullus for the production of poetry.

By taking into consideration all such intertext candidates in *Angustae Vitae*, we will show that machine classification is able to produce statistically strong validation results. We present our study of the three style markers in this context, highlighting strengths and weaknesses. We hope that this case study will serve as a first step towards a more sophisticated and efficient analysis of intertextuality in general. Moreover, this work raises important linguistic questions on the nature of conscious and subconscious influence in style, which is an area we intend to explore in further work.

References

- Diederich, J., Kindermann, J., Leopold, E. and Paass G.** (2003). 'Authorship Attribution with Support Vector Machines'. *Applied Intelligence*. **19(1-2)**: 109-123.
- Forstall, C.W. and Scheirer, W.J.** (2009). 'Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound'. *Chicago Colloquium on Digital Humanities and Computer Science*. Chicago.
- Holmes, D., Robertson, M., and Paez, R.** (2001). 'Stephen Crane and the New York Tribune: A Case Study in Traditional and Non-traditional Authorship Attribution'. *Computers and the Humanities*. **35(3)**: 315—331.
- Kristeva, J.** (1986). 'Word, Dialogue and Novel'. *The Kristeva Reader*. Moi, T. (ed.). New York: Columbia University Press, pp. 34—61.
- Paul the Deacon** (1881). 'Carmina'. *Monumenta Germaniae Hisotica, Poeta Latini Aevi Carolini Vol 1*. Diemmler, E. (ed.). Berlin.

Historical Interpretation through Multiple Markup: The Case of Horatio Nelson Taft's Diary, 1861-62

Garfinkel, Susan

sgarfinkel@loc.gov

Library of Congress, Washington DC, USA

Heckscher, Jurrett Jordan

jhec@loc.gov

Library of Congress, Washington DC, USA

1. Background

Now fifteen years old, the still-growing American Memory Web site at the Library of Congress (<http://memory.loc.gov/>) offers more than 130 separate and diverse multi-media collections — comprising more than a million digitized library items — to a vast base of virtual patrons across the Internet. Yet despite radical changes in the ways that people have come to use the World Wide Web, the fundamental conception of American Memory and its collections remains much as envisaged fifteen years ago. What happens next to such established but largely static digital resources? Aside from implementing obvious upgrades such as higher-resolution image scans, cleaned-up OCR, fleshed-out metadata, or faceted search capabilities, how might cultural repositories more fundamentally enhance their existing online content? Can we fully (re)imagine a second iteration, a next generation, for already digitized historical materials?

Our demonstration project works to consider this question for a single American Memory collection, purposely setting aside issues of scale or interoperability in favor of exploring effective and compelling ways to convey to users the particular character of specific historical resources. Starting with the unpublished manuscript diary of Horatio Nelson Taft (1861-62), already digitized (as textual transcription with page images) at <http://memory.loc.gov/ammem/taft/html/tafthome.html>, we have chosen to explore the implications of creating several alternate and explicitly interpretive frameworks by applying multiple XML markup to this bounded but compellingly dense and historically significant text.

American Memory patrons bring unusually diverse research needs to the same body of historical materials: typical users include everyone from elementary school students and hobbyists to lawyers, librarians, college professors and members

of Congress. Our project's concerns stem from this diversity as well as from our own role as scholars of American culture working with historical materials in a library setting. Traditional library practice treats sources as analytically separate from their interpretation, even while the most basic interventions of librarianship are fundamentally interpretive. Historians, by contrast, well recognize their own work with sources as interpretive, but still tend to view textual sources as fully fixed in meaning.

Textual analysis is an underdeveloped tool in American historical study, and history as a discipline has lagged behind literature in imagining computing as a threshold for innovative research. Historians' traditional use of computing has trended quantitative rather than qualitative, while recent innovations emphasize creating tools to assist the mechanics of research or scholarly interaction rather than transforming them. Explorations of how computing might fundamentally change the practices of history or the outcomes of historical interpretation are still needed.

Against this background the Taft Diary project takes on several related goals: (1) to visibly demonstrate a historical text's accessibility to multiple simultaneous interrogations enacted through digital scholarship; (2) to more fully explore the multidimensionality of a significant historical text; (3) to introduce the methods and questions of digital humanists more centrally into a library context (making them better known to library practitioners and more widely available to diverse library audiences); and (4) to meld the questions and methods of historians with the advances in digital textual scholarship arising among literary scholars. With our project we hope to establish a model for text-based scholarship — literary and ethnographic as well as historical — that foregrounds the individual user's interpretative needs while also conveying that interpretation to the broad variety of the diary's potential users.

2. The artifact and its text

The Horatio Nelson Taft Diary commands wide historical interest. When it was introduced to the public on the Library's Web site in 2001, it offered the first new information about the events of Abraham Lincoln's death to come to light in half a century. In fact the diary emerges from the confluence of multiple transformative historical developments: not only the American Civil War and Lincoln's presidency, but profound long-term changes including the rise of the American middle class; the nation's industrial, technological, and consumer revolutions; the decisive expansion in the size and influence of the federal government; and

the maturation of Washington, DC, as a complex urban community of national and international importance.

An exceptional historical record on many levels, the Taft Diary is also unusually well suited to an experiment in simultaneous multidimensional interpretive markup. Its contents are rich in the overlap of ordinary life and significant events, people, and places, thereby appealing both to specialized historians and to a broad general audience. Its limited size offers definite boundaries to the range, though not the depth, of interpretive markup. Moreover, the formulaic pattern underlying most of its daily entries invites ready comparison while ensuring an organizing degree of structural regularity.

Taft's three-volume manuscript diary consists of daily entries from January 1, 1861, through April 11, 1862 (the end of vol. 1), and less frequent entries through May 30, 1865. The current phase of our project deals only with vol. 1, which contains in total 466 entries. Because Taft used a printed blank diary book for vol. 1, each 1861 entry is eleven handwritten lines long, or typically ninety to one hundred words (fig. 1). (The volume's 1862 entries, written into pages intended for back matter, vary more in length.) Beyond the consistency of the entries' length, certain content elements recur so frequently as to be almost formulaic: information about the weather, illnesses within and beyond the family circle, political and military news about the Civil War, and the people and locations that populate Taft's daily life.



Figure 1: Taft Diary vol. 1, entries for Sept. 19-21, 1861; transcription of Sept. 20 entry

3. Implementation

In selecting markup schemas, we considered and discarded a number of interpretive typologies, including Library of Congress Subject Headings (LCSH) (<http://www.loc.gov/aba/cataloging/subject/>) and the Library of Congress Classification Outline (<http://www.loc.gov/catdir/cpsoc/lcco/>), because they turn out to lack a consistent level of granularity across cultural and historical domains. Seeking to demonstrate applicability and variety, for phase one we selected three schemas for their breadth as well as depth, their flexibility, and their inherent relevance to the content of our text:

1. TEI, <http://www.tei-c.org/>
2. Outline of Cultural Materials (OCM), <http://www.yale.edu/hraf/outline.htm>
3. Scholar-defined "webs of significance" such as:
 - i. Taft's recurring concerns
 - a. weather
 - b. places
 - c. illness
 - ii. networks
 - a. of social relationships
 - b. how news is conveyed

These three schemas together provide for theoretical and well as methodological diversity, representing a variety of approaches to the analysis of historical texts. TEI is a widely used standard for the representation of texts in digital form, with well-developed editorial standards and an established community of users. The OCM is anthropological, deriving from attempts to develop comprehensive categorizations of human cultural phenomena. Scholar-defined "webs of significance" follow from our own close readings of the text. In the future we anticipate exploring additional markup typologie based in the Thesaurus for Graphic Materials (<http://www.loc.gov/pictures/collection/tgm/>), GIS, and data visualizations, so as to demonstrate the diversity and versatility of interpretation that simultaneous multiple markup can sustain.

Our baseline text for markup is the transcription of Taft's vol. 1, which was originally provided in SGML mapped to the American Memory DTD. Our own markup reverses normalizations in transcription in favor of a diplomatic version, and strips out the older SGML tags.

Implementation raises methodological and technical challenges. In our process we must move from

theoretical issues to the development of standards for each tag category, exploring the dilemmas that intensive and self-created markup entails: for example, whether phrases or sentences should be the unit of markup with OCM tags (fig. 2). Resolving such questions requires us to confront the practical challenges of creating markup conventions originally based in theory rather than established practice. Theoretical language about multiplicity is inspiring — and even starts to get at the truth of lived experience — but in practice we must also make sound choices about how to complete the markup with reasonable consistency so that others may use and rely on it. Further, do we mark the same text multiple times, or invest ourselves in some version of a standoff markup (see the XStandoff toolkit, for example)? How do we best present multiple interpretive options to our audience? At this level, the Taft Diary project is still a work in progress.

FRIAD 20	(September 20, 1991)					
756					421	
727					421	
476	743		328	324	476	471
<All the <Indiana Hospital (Pharmacy)> today, 100 Patients/>	<Black Rock> for <France> <Get a <Cloud from the Land Office>>					
467	121	188	541	583	724	367
<Land Office>> for the Police of <Lyons>>	<Attended the <Parade> with <Julia>> of the <Regiment> on <July 2000>					
374		363		554		708
Square</> <Called with her upon the Woodbury>>	<129 SE> -> <met all educated people> and much					
758						
727						
732	438				704	
devoted to the <Rock audience>>	<People are pouring in rapidly now, from five to six <Regiments> per day					
403		9027	805	726		
<Refugee by the Danube>>	Large numbers have gone over the river>	<Big week>>	<An attack on the City			
expected>						
key:						
131 Location		553 Talent Mobility				
138 Cultural Identity and Pride		578 Visiting and Hospitality				
238 Carpentry		583 Family Relationships				
263 Streets and Traffic		647 Administrative Agencies				
387 Parks		676 Contracts				
423 Real Property		679 Agency				
425 Acquisition and Retirement of Property		724 Ground Combat Forces				
446 Military Aid		725 Security Corps				
448 Total		726 Warfront				
461 Vehicles		727 Aftermath of Conflict				
505 Water Transport		743 Hospitals and Clinics				
534 Musical Instruments		752 Disability Workers				
541 Speculations		758 Mental Care				
554 Status, Role and Prestige		818 Corrections of Time				

Figure 2: Text of Taft Diary vol. 1, Sept. 20 entry showing conceptual phrase-level OCM markup

4. Conclusion

We anticipate that simultaneous multiple markup will render the text dynamic, surfacing its performative and ritualized aspects; and that it will invite new attention to its literary and linguistic aspects. Markup can also foreground key aspects of worldview of which the writer himself was unaware. In these and other ways, simultaneous multiple markup becomes a new instrument of historical interpretation, reimagining analysis beyond the realm of narrative prose.

While our technical questions are not yet fully answered, and the project itself is ongoing, we seek

to share important methodological and theoretical questions, and our conclusions so far, with our colleagues in digital humanities.

References

- Cohen, Daniel J., Rosenzweig, Roy** (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/>.

Geertz, Clifford (1973). 'Thick Description'. *The Interpretation of Cultures: Selected Essays*. New York: Basic Books, pp. 3-30.

McGann, Jerome (2004). 'Marking Texts of Many Dimensions'. *A Companion to Digital Humanities*. Schreibman, Susan, Siemens, Ray, Unsworth, John (eds.). Oxford: Blackwell, pp. 198-217. <http://www.digitalhumanities.org/companion/>.

Milligan, Frank D. (2007). 'A City in Crisis: The Wartime Diaries of Horatio Nelson Taft'. *President Lincoln's Cottage*. **entry of August 1**, blog. <http://lincolncottage.wordpress.com/2007/08/01/a-city-in-crisis-the-wartime-diaries-of-horatio-nelson-taft/>.

Sellers, John (February 2002). 'Washington in Crisis, 1861-1865: Library Acquires the Diary of Horatio Nelson Taft'. *Library of Congress Information Bulletin* **61, no. 2**. <http://www.loc.gov/loc/lcib/0202/cw-diarist.html>.

Stührenberg, Maik, Jettka, Daniel (2009). 'A Toolkit for Multi-Dimensional Markup: The Development of SGF to XStandoff'. *Proceedings of Balisage: The Markup Conference 2009*. Balisage Series on Markup Technologies. 3 vols. <http://www.balisage.net/Proceedings/vol3/html/Stuhrenberg01/BalisageVol3-Stuhrenberg01.html>.

Diple, modular methodology and tools for heterogeneous TEI corpora

Glorieux, Frédéric

frederic.glorieux@enc.sorbonne.fr

École nationale des chartes

Canteaut, Olivier

olivier.canteaut@enc.sorbonne.fr

École nationale des chartes

Jolivet, Vincent

vincent.jolivet@enc.sorbonne.fr

École nationale des chartes

The *École nationale des chartes* publishes a variety of electronic corpora,¹ focused on historical sources (medieval, but also, modern and contemporary). A dictionary,² a collection of acts,³ or a manuscript⁴ are very different types of documents, each requiring different structures and interfaces. A narrative manuscript needs a table of contents, a dictionary, fast access to headwords, and acts, the ability to sort by dates. Each editorial project should allow customization, but efficient development requires that the tools and corpora are as normalized as possible. New needs emerge, such as natural language processing research, requiring large corpora with normalized metadata sets and word tagging. For several months we have been working on a platform to address these needs: *Diple* is a collection of tools to organize modular production, publication, and searching of electronic corpora.

1. Modular schemas

The TEI guidelines (500 XML elements, 1500 pages of documentation) allow endless variations in encoding, even for identical objects. For example, *italic* in our corpora has been encoded with different combinations of <(hi|emph) rend="(italique|italic|ital.|i|itlaic|...)">. After several years of development, with different encoders, each electronic edition becomes an independent software, with its own encoding, mistakes, workarounds, with also different technologies for publication or fulltext searching.

Diple starts with housekeeping. First, for all our tagged texts, we wrote a precise *document type definition* (in Relax-NG syntax) in order to define three main and shared schemas :

- file metadatas (<teiHeader>)
- general text (blocks and inlines)

- structure for a specific type of text (ex: acts and charters, dictionaries)

The normalization of the corpora is a more sustainable investment than new software. These shared schemas are extremely helpful for normalizing and validating XML instances, and therefore allow us to take advantage of earlier TEI editions. Of course, the *Diple* TEI schema is modular, allowing customization for each editorial project.⁵ An editor can then focus on the specificities of each edition. Are named entities sufficiently tagged to generate automatic indexes? Are the sentences chunked, the words lemmatized?

Moreover, this work of normalization of our XML corpora is a small price to pay to factorize our code, for instance to create a standard XSLT engine: the screen transformation of a new corpus conforming to those schemas is done by this engine, increasing our publication productivity. In the end, the XSLT of a specific edition is short, focusing on the very specific aspects of the corpus (its custom schema) related to a research project, the main part of the publication job being done by the *Diple* XSLT engine. The same logic applied for presentation CSS.

2. Shared interface components, documents driven

A publication system usually allows templating and plugins. A good software architecture should be conceived in this way, but scholarly editions don't function like a CMS. Templating systems are usually designed to effect easy change of colours, to deliver the same feature under different designs. In a scholarly collection, books could share a cover, but follow very different structures. Rather than constrain all corpora to a single template, the *Diple* system provides different components, allowing an electronic corpus editor to compose the interface best suited to his text. Headers or footers are easy to share, but beyond that, one project might require a fulltext search box, another a database query, another a sidebar table of contents. Design snippets or plugins are conceived of as portal bricks, easy to compose in a server page (PHP), and are kept as simple and light as possible. If a local variable, function or object could have a general interest, it should be shared.

3. Text engines for research, retrieving and concordances

Navigation, tables and indexes, should answer most of the user's needs; but a search box is also an important navigation tool. *Diple* ensures a canonical electronic publication, with persistent addresses, so that different text engines can be plugged around the edition. Corpora may require

different approaches. A collection of items, like a dictionary or cartularies, needs at first a retrieving engine to get an item conveniently, by a keyword in full-text, but also dates, headwords and other metadatas. There are also texts for which no divisions are relevant; a concordance report is much more informative, displaying all occurrences in context. Different tools offer different views, documented XML allows us to generate what an engine likes. We have successfully used MySQL full-text indexes⁶ for navigation interfaces, PhiloLogic⁷ for concordances, Lucene⁸ is very efficient to retrieve items, and we learned to use IMS Corpus Workbench (CWB)⁹ for future lemmatized corpora. But sometimes we also simply use mixed scripts (shell, XSLT, SAX...) to run a specific experiment on a word or a semantic field.

4. Conclusion

Diple grows and adapts with each new corpus, rapidly incorporating other corpora, an idea worth generalizing. All our code will soon be released under a free software license. Anyone can download, read, and try *Diple*. We don't claim it will work for all your TEI documents, but if they conform to the schemas, you will quickly get nice results on the screen.

References

- Bourgain, Pascale, Vieillard, Françoise (coord.)** (2002). *Conseils pour l'édition des textes médiévaux*. Fascicule III. Textes littéraires. Paris: Éd. du CTHS, École des chartes.
- Guyotjeannin, Olivier, Vieillard, Françoise (coord.)** (2001). *Conseils pour l'édition des textes médiévaux*. 'Conseils généraux'. Paris: Éd. du CTHS, École des chartes.
- Olivier Guyotjeannin (coord.)** (2001). *Conseils pour l'édition des textes médiévaux*. 'Actes et documents d'archives'. Paris: Éd. du CTHS, École des chartes.
- McCandless, Michael, Hatcher, Erik, Gospodnetić, Otis** (2010). *Lucene in action*. Manning Publications Co., 2e ed.. <http://www.manning.com/hatcher3/>.
- TEI Consortium (ed.). TEI P5: Guidelines for Electronic Text Encoding and Interchange.** <http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/REF-ELEMENTS.html>.
- Wooldridge, Russon** (1997). *Les Débuts de la lexicographie française*. Toronto: EDICTA2e éd.. <http://www.chass.utoronto.ca/~wulfric/edicta/wooldridge/>.

Notes

1. <http://elec.enc.sorbonne.fr/>
2. <http://ducange.enc.sorbonne.fr/>
3. <http://elec.enc.sorbonne.fr/cartulaires/>
4. <http://elec.enc.sorbonne.fr/sanctoral/>
5. For example, <http://elec.enc.sorbonne.fr/cartularesIdF/src/schema.htm>
6. <http://dev.mysql.com/doc/refman/5.0/en/fulltext-search.html>
7. <http://philologic.uchicago.edu/>
8. <http://lucene.apache.org/java/>
9. <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>

A New Spatial Analysis of the Early Chesapeake Architecture

Graham, Wayne

wayne.graham@virginia.edu

University of Virginia

Practitioners of the "new social history," which came to prominence beginning in the 1960s and 70s, utilized digital tools and data-driven methodologies to glean an understanding of people who left little documentary record of their daily lives. Perhaps most enduring of these techniques has been the utilization of quantitative methods to describe communities and to build better arguments about the daily lives of historical subjects. The focus for these historians is not only the high points thought worthy of record in diaries, newspapers, or court papers, but also how quotidian interaction and daily chores – such as cooking, cleaning, or plowing – were accomplished and how their patterns differed across regions. While quantitative techniques have yielded a rich analysis of the past, temporal, social, and geographic dimensions of historical data often diverge and can be muddled in the choices scholars make about how best to tell their story given time and resource considerations, as well as how to argue the larger points of a particular person's, event's, or object's societal influence.

In our recent work at the Scholars' Lab at the University of Virginia Library (where we support geospatial technology in the humanities and social sciences and have recently played host to an NEH-funded Institute for Enabling Geospatial Scholarship and a Mellon Scholarly Communication Institute on spatial tools and methods), we have been advocating the idea that incorporating geographic information systems into projects can yield interesting new interpretative apparatus for scholarship. This is neither a new concept, or an especially easy path to take. Martyn Jessop has detailed the obstacles to incorporation of geospatial information in humanities research in the pages of *Literary and Linguistic Computing*.¹ However, to test the approaches we advocate to others, I decided to revisit a project that I undertook a few years ago with several prominent historians and archaeologists of the architectural development of the Colonial Chesapeake.² While the data resulting from their work is the basis of two important essays on Chesapeake architecture, and additionally served as the framework for an NEH grant investigating the development of slave quarters in Virginia, it has

languished and few outside the project team actually know of the data's existence.³ I considered this a perfect example of an important project to rethink by adding a more defined geographic dimension to its analytical approach. Could the application of GIS technologies further test our long-held beliefs about the development of the Chesapeake?

In their seminal essay on "impermanent" Chesapeake architecture, Cary Carson, Norman Barka, William Kelso, Gary Wheeler Stone, and Dell Upton first attempted systematically to synthesize and analyze data extracted from several investigations into early Chesapeake architecture.⁴ This article was squarely focused on the structures settlers built between first shelters and more durable buildings. Despite the genius of their work, the Carson team was limited in that archaeological work in the Chesapeake region was then still young, and the data from a scant two dozen sites supported their analysis.

In the nearly three decades since this piece was published, more than ten times that number of sites has been identified and excavated. However, this boom in investigation of the Colonial Chesapeake resulted not in masses of usable data for broad-scale analysis, but in the explosion of a so-called "gray literature"—reports produced for project clients and funding organizations, but circulated only in limited numbers. Often, after their initial compilation, these reports have languished in state or institutional archives and little systematic work has been done to organize, or even make available, this often tangled mass of data. As a result, the accumulation of archaeological data has far outpaced its published analysis. Further complicating matters are embargoes placed upon research reports (usually meant to help protect against artifact theft) that even further distance access to raw facts on these early sites from the hands of researchers.

In conjunction with celebrations marking the 400th anniversary of the founding of the Jamestown settlement, a new team (consisting of Willie Graham, Carter Hudgins, Carl Lounsbury, Fraser Neiman, James Whittenburg, and myself) looked to the more recent archaeology.⁵ Having collected references to archaeological sites mentioned in articles, research reports, conference proceedings, and in personal interviews, Willie Graham of the Colonial Williamsburg Foundation amassed an index of known sites dated before ca. 1720 in the Colonial Chesapeake. From this index, our research team designed a data model that provided a crucial new dimension into this particular facet of history by combining solid statistics pertaining to material culture with an appreciation for the historical discourse in this area of study. Dubbed the Database of Early Chesapeake Architecture

(DECA), we took a quantitative approach to this expanded set of archaeological and architectural data making it possible for the first time to accurately date significant shifts in the cultural repertoires of Chesapeake colonists and link them in convincing – and testable – ways to the unique ecological, economic, and social conditions to which they were a response. Through the use of solid data modeling techniques, information from hundreds of new archaeological and architectural investigations provided a fresh opportunity to analyze the emergence of regional building practices and chart the dynamics of social interaction in the tobacco colonies through the arrangement of planters' houses and outhouses, as well as in the types of goods the colonists possessed and food they consumed.

When the database was initially designed, it was composed of a handful of simple tables detailing building and phase dates, dimensions, floor plan types, chimney types, and foundation characteristics and documented using the unified modeling language (UML). As the project progressed, the database structure evolved to include owner information and documentary references, resulting in a complex implementation of relational tables. However, the only documentation of place was in the recording of a town or county in which the site was located.

My current work reimagines the original DECA project to include not only its core statistical information, but also well-defined geographic locations allowing scholars to ask new questions of the data and visualize them in new and compelling ways. Through the addition of well-constructed geospatial information, and the application of tools and methods, we are refining a more striking analysis of the Chesapeake data for the use of our faculty collaborators in the Scholars' Lab. A new presentation, not only of traditional statistical outputs (distribution curves, ANOVA tables, etc.), but of distribution patterns in architectural and archaeological details manifested across time and across the landscape of the Chesapeake, affords researchers even more insight into regional differentiation in building patterns, and more striking opportunities to display and engage their data. This presentation will describe the spatial tools and methods we advocate in the Scholars' Lab including the use of the PostGIS data store, Ruby on Rails (with the GeoKit gem), and OpenLayers, outline their application to the Chesapeake dataset, and offer some observations on lessons (both methodological and substantive) learned in my revisiting of this digital humanities project through the lens of geospatial analysis.

Notes

1. Martyn Jessop, "The Inhibition of Geographical Information in Digital Humanities Scholarship," *Lit Linguist Computing* (November 20, 2007): fqm041.
2. Willie Graham, Carter L. Hudgins, Carl R. Lounsbury, Fraser D. Neiman and James P. Whittenburg, "Adaptation and Innovation: Archaeological and Architectural Perspectives on the Seventeenth-Century Chesapeake," *The William and Mary Quarterly* 64, no. 3 (July 2007), <http://www.historycooperative.org/journals/wm/64.3/graham.html>
3. See Doug Sanford (University of Mary Washington) and Dennis Pogue (Mount Vernon), "Measuring the Social, Spatial, and Temporal Dimensions of Virginia Slave Housing," National Endowment for the Humanities, 2009 and Cary Carson et al., "New World, Real World: Improvising English Culture in Seventeenth-Century Virginia," *Journal of Southern History* LXXIV (February 2008).
4. Cary Carson et al., "Impermanent Architecture in Southern American Colonies," in *Material Life in America, 1600-1860* (Boston: Northern University Press, 1988), 113-158.
5. The data that the "Adaptation and Innovation: Archaeological and Architectural Perspectives on the Seventeenth-Century Chesapeake" and "New World, Real World: Improvising English Culture in Seventeenth-Century Virginia," articles were based on is available for browsing at <http://deca.swem.wm.edu>

The Importance of Pedagogy: Towards a Companion to Teaching Digital Humanities

Hirsch, Brett D.

brett.hirsch@gmail.com

University of Western Australia

Timney, Meagan

mbtimney.etcl@gmail.com

University of Victoria

The need to “encourage digital scholarship” was one of eight key recommendations in *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences* (Unsworth et al). As the report suggested, “if more than a few are to pioneer new digital pathways, more formal venues and opportunities for training and encouragement are needed” (34). In other words, human infrastructure is as crucial as cyberinfrastructure for the future of scholarship in the humanities and social sciences. While the Commission’s recommendation pertains to the training of faculty and early career researchers, we argue that the need extends to graduate and undergraduate students. Despite the importance of pedagogy to the development and long-term sustainability of digital humanities, as yet very little critical literature has been published. Both the *Companion to Digital Humanities* (2004) and the *Companion to Digital Literary Studies* (2007), seminal reference works in their own right, focus primarily on the theories, principles, and research practices associated with digital humanities, and not pedagogical issues. There is much work to be done.

This poster presentation will begin by contextualizing the need for a critical discussion of pedagogical issues associated with digital humanities. This discussion will be framed by a brief survey of existing undergraduate and graduate programs and courses in digital humanities (or with a digital humanities component), drawing on the “institutional models” outlined by McCarty and Kirschenbaum (2003). The growth in the number of undergraduate and graduate programs and courses offered reflects both an increasing desire on the part of students to learn about sorts of “transferable skills” and “applied computing” that digital humanities offers (Jessop 2005), and the desire of practitioners to consolidate and validate their research and methods. We propose a volume, *Teaching Digital Humanities: Principles,*

Practices, and Politics, to capitalize on the growing prominence of digital humanities within university curricula and infrastructure, as well as in the broader professional community.

We plan to structure the volume according to the four critical questions educators should consider as emphasized recently by Mary Bruenig, namely:

- What knowledge is of most worth?
- By what means shall we determine what we teach?
- In what ways shall we teach it?
- Toward what purpose?

In addition to these questions, we are mindful of Henry A. Giroux’s argument that “to invoke the importance of pedagogy is to raise questions not simply about how students learn but also about how educators (in the broad sense of the term) construct the ideological and political positions from which they speak” (45). Consequently, we will encourage submissions to the volume that address these wider concerns.

References

- Breunig, Mary** (2006). 'Radical Pedagogy as Praxis'. Radical Pedagogy. http://radicalpedagogy.icaap.org/content/issue8_1/breunig.html.
- Giroux, Henry A.** (1994). 'Rethinking the Boundaries of Educational Discourse: Modernism, Postmodernism, and Feminism'. *Margins in the Classroom: Teaching Literature*. Myrsiades, Kostas, Myrsiades, Linda S. (eds.). Minneapolis: University of Minnesota Press, pp. 1-51.
- Schreibman, Susan, Siemens, Ray, Unsworth, John (eds.)** (2004). *A Companion to Digital Humanities*. Malden: Blackwell.
- Jessop, Martyn** (2005). 'Teaching, Learning and Research in Final Year Humanities Computing Student Projects'. *Literary and Linguistic Computing*. **20.3 (2005)**: 295-311.
- McCarty, Willard, Kirschenbaum , Matthew** (2003). 'Institutional Models for Humanities Computing'. *Literary and Linguistic Computing*. **18.4 (2003)**: 465-89.
- Unsworth et al.** (2006). *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. New York: American Council of Learned Societies.

A Bilingual Digital Edition of Trinity College Cambridge MS O.1.77.

Honkapohja, Alpo

alpo.honkapohja@helsinki.fi

University of Helsinki

The poster will present my work-in-progress PhD project of a 15th-century bilingual medical manuscript, containing Latin and Middle English. The edition is designed with the needs of historical linguistics in mind, and will have some corpus functionalities. My long term aim is to use it as a pilot study of sorts in contrastive investigation of Latin and Middle English medical writing.

1. Background

Medieval medical writing for a long period of time received fairly little attention. For instance, Robbins described it, in 1970, as a “Yukon territory crying out for exploration”. In the 1990s and 2000s, the situation has changed, and the field is becoming filled with tiny flags stating the claims of various research projects and individual scholars. There are now large electronic corpora such as the *Middle English Medical Texts* (MEMT), published 2005, and *A Corpus of Middle English Scientific Prose*, currently being compiled in collaboration between the University of Malaga and Hunter Library in Glasgow.

These resources do, however, have one inherent bias. They focus on Middle English material, which gives a distorted view of the linguistic situation in England in the late Middle Ages. England, after the Norman conquest, was a trilingual society in which educated members of the society were likely to have at least some degree of literacy in Latin, Anglo-Norman French as well as English. This shows, for instance, in the fact that manuscripts containing texts in more than one language outnumber monolingual ones. (cf. Voigts 1989). Moreover, marginal comments also suggest they had a readership competent in more than one language.

My PhD project is intended as the first genuinely bilingual online resource of medical manuscripts in late Medieval England, and will hopefully pave the way for similar resources in the future. It is designed for both historical linguists and historians, but paying special attention to the needs of linguistics.

2. Trinity College Cambridge, MS O.1.77.

Trinity MS O.1.77. is a pocket-sized (75 x 100 mm) medical handbook, located in Trinity College Cambridge. It contains 10 to 18 texts on medicine, astrology and alchemy. It is usually treated as a sibling MS of the so-called Sloane-group of Middle English manuscripts, which is a group of late Latin, English and French MSS originating from London or Westminster in the late Middle English period (cf. e.g. Voigts 1990). James assigns MS Trinity O.1.77 an exact date 1460, based on astrological markings in the final flyleaf (1902), although it may not be entirely accurate. (see Honkapohja 2010, forthcoming)

Roughly 4/5 of the manuscript is in Latin and 1/5 in English, that is, out of slightly less than 30,000 words, c. 24,000 words are Latin and 5,500 in English. There does not appear to be a clear-cut division between prestigious Latin texts and more popular English ones. Latin, however, is used almost exclusively for metatextual functions such as incipits and explicits. Nearly all marginal comments in the manuscript are in Latin.

3. The digital edition

The digital edition which I am preparing will be designed in such a way that it will function as reliable data for historical linguistics. This involves encoding a sufficient amount of detail on linguistic variants without normalising, modernising, or emending the data, and keeping all editorial interference transparent (see e.g. Kytö, M., Grund P. and Walker T. 2007 or Lass 2004)

On the technical side, I am using TEI P5 –conformant XML tagging built on stand-off architecture. Things included in the base-level annotation are a graphemic transcription of the text (cf. e.g. Fenton & Duggan 2006), select manuscript features such as layout, and information about the manuscript and hand. Each word will also be tagged with a normalised form, useful for linguistic research, and an ID which allows the addition of additional tagging by means of stand-off annotation – including, for instance, POS tagging, semantic annotation or lemmatisation.

The edition will have an online user interface, which will allow the user to select the level of detail he or she wishes. It will be possible to use it with either normalised text or diplomatic transcription. It will be released under a Creative Commons license. The user will have full access to the XML-code, including all levels of annotation, and will be allowed to download and modify it for non-commercial purposes.

4. DECL

The development of the edition will take place in collaboration between the Digital Editions for Corpus Linguistics (DECL) project based at the University of Helsinki.

The DECL project was started by three post-graduate students in 2007. It aims to create a framework for producing online editions of historical manuscripts suitable for both corpus linguistic and historical research. DECL editions use a more strictly defined subset of the TEI-guidelines and are designed especially to meet the needs of corpus linguistics. The framework consists of encoding guidelines compliant with TEI XML P5. The aims of the project are presented in more detail in our article (Honkapohja, Kaislaniemi & Marttila 2009).

5. Digital Edition of O.1.77 as a resource for the study of bilingualism

My PhD project has both short and long term goals related to the study of multilingualism. The short term aim is to design the edition in a way that is of maximum use for scholars working with medical texts and especially multilingualism. I am especially putting a lot of effort into interoperability and making the encoding as flexible as possible.

Hypothetical research questions for the edition will include, for instance:

- *Spelling variation.* Using the edition will enable getting information on spelling variation in English and Latin, in order to see whether the accepted general view that Latin was more regular is supported by quantitative data.
- *The use of brevigraphs and contracted forms.* Manuscript abbreviations are an extremely common feature in the Latin texts of the manuscript. They are also applied in the Middle English sections, but with less frequency. The edition will make it possible to obtain exact statistical information on which manuscript abbreviations carry into the vernacular, and with how much variation and frequency.
- *Syntactic complexity:* Do sentences in Latin contain a greater number of sub clauses and other signs of syntactic complexity than Middle English ones?
- *Textual Functions:* The use of English and Latin in various text types, recipes, metatextual passages (in which Latin very much dominates). The type of structural and background information which is being annotated in the edition will enable the user to perform the searches on different level

textual passages, including marginal comments and metatextual passages.

After the completion of the PhD project, the edition will be expanded with other related multilingual medical and alchemical manuscripts in the Sloane group, which will increase the usefulness of the database, by allowing, for instance, comparative study of the same text in different manuscripts. I am also planning to make use of the available corpora on Middle English medical writing for comparisons to Middle English.

References

- A Corpus of Middle English Scientific Prose.* <http://hunter.filosofia.uma.es/manuscripts> (accessed 13 March 2010).
- Fenton, E. G., Duggan, H. N.** (2006). 'Effective Methods of Producing Machine-readable Text from Manuscript and Print Sources'. *Electronic Textual Editing*. Burnard, L., O'Brien O'Keeffe, K., Unsworth, J. (eds.). New York: MLA.
- Honkapohja, A., Kaislaniemi, S., Marttila, V.** (2009). 'Digital Editions for Corpus Linguistics: Representing manuscript reality in electronic corpora'. *Corpora: Pragmatic and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)*. Ascona, Switzerland, 14-18 May 2008. Jucker, A. H., Schreier, D., Hundt, M. (eds.). Amsterdam/New York: Rodopi.
- Honkapohja, A.** (2010: forthcoming). 'Multilingualism in Trinity College Cambridge Manuscript O.1.77'. *Studia Anglica Posnaniensia*.
- James, M. R.** (1902). *The Western Manuscripts in the Library of Trinity College, Cambridge. A Descriptive Catalogue*. Cambridge: CUP V. III, Containing an Account of the Manuscripts Standing in Class O. .
- Lass, R.** (2004). 'Ut custodiant litteras: Editions, Corpora and Witnesshood'. *Methods and Data in English Historical Dialectology*. Dossena, M., Lass, R. (eds.). Linguistic Insights. Bern: Peter Lang.
- Robbins, R.H.** (1970). 'Medical Manuscripts in Middle English'. *Speculum*. No.3 Jul 1970: 393-415. <http://www.jstor.org/stable/2853500> (accessed 13 March 2010).
- Taavitsainen, I., Pahta, P., Mäkinen, M. (eds.)** (2005). *Middle English Medical Texts*. Amsterdam: John Benjamins, CD-ROM.
- Text Encoding Initiative (TEI)*. <http://www.tei-c.org> (accessed 13 March 2010).

Voigts, L. E. (1989). 'Scientific and Medical Books'. *Book Production and Publishing in Britain 1375-1475*. Griffiths, J., Pearsall, D. (eds.). Cambridge: Cambridge University Press.

Voigts, L. E. (1990). "The "Sloane Group": Related scientific and medical manuscripts from the fifteenth century in the Sloane Collection". *The British Library Journal*. **16**: 26-57.

Kytö, M., Grund P., Walker T. (2007). 'Regional variation and the language of English witness depositions 1560-1760: constructing a 'linguistic' edition in electronic form'. Pahta, P., Taavitsainen, I., Nevalainen, T., Tyrkkö, J. (eds.). *Towards Multimedia in Corpus Studies. Studies in Variation Contacts and Change in English*. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki . http://www.helsinki.fi/varieng/journal/volumes/02/kyto_et_al (accessed 13 March 2010).

The Craig Zeta Spreadsheet

Hoover, David L.

david.hoover@nyu.edu
New York University, USA

Zeta, a new measure of textual difference introduced by John F. Burrows, can be used effectively in authorship attribution and stylistic studies to locate an author's characteristic vocabulary—"marker" words that one author uses consistently, but another author or authors use much less frequently or not at all (Burrows 2005, 2006; Hoover 2007a, 2007b, 2008). Zeta analysis excludes the extremely common words that have traditionally been the focus and concentrates on the middle of the word frequency spectrum. Beginning in 2008, I developed an Excel spreadsheet implementation of Zeta and its related measure Iota (which focuses on words that are relatively rare in one author, but extremely rare or absent from others), available on my Excel Text-Analysis Pages. Recently, however, Craig has developed an alternative version of Zeta that simultaneously creates sets of marker words and anti-marker words and has applied it impressively to Shakespeare authorship problems (Craig and Kinney, 2009; Hoover, forthcoming). Although Craig's Zeta focuses on the same part of the word frequency spectrum as Zeta, its calculation and results are quite different. Because it seems poised to become an important tool for computational stylistics, I have created and will demonstrate the Craig Zeta Excel spreadsheet that automates its calculation.

Craig's Zeta is a powerful but simple method of measuring differences among authors. It begins with two sets of texts divided into about equal-sized sections, then compares how many sections for each author contain each word, ignoring the frequencies of the words and concentrating on their consistency of appearance. The most natural comparison is between two authors, but it can be used to study any contrast. My demo version contrasts thirteen female and thirteen male American poets born between 1911 and 1943 (about 8,000 words of poetry by each, divided into two sections).

The heart of the method is that it combines the ratio of the sections by one author in which each word occurs with the ratio of the sections by the other author from which it absent into a single measure of distinctiveness for each word. Zeta scores theoretically range from two (for a word found in every section by one author and absent from every section by the other), to zero (for a word found in

no sections by one author and in all sections by the other). Sorting the words on this composite score produces two lists of words, one favored by the first author and avoided by the second, the other favored by the second author and avoided by the first.

The snippet from the spreadsheet in Fig. 1 (shown before the macro operates) will clarify the calculation. In E7 and E8, the user enters labels (automatically copied into columns A and G and Row 9) for the two groups to be compared. The data to be analyzed is in columns H through CA, rows 11ff, with most columns minimized so the various categories are visible. The combined word frequency list for the two groups is in column G, with the raw frequencies for each word in each section in columns H-CA. The calculation is performed in columns A-E. Column D sums the sections of poetry by women that contain the word, and column E sums the sections of poetry by men that do not contain the word. The most frequent words typically occur in all sections, but note that *me*, the 30th most frequent word, is absent from one of the men's sections. Column B calculates the ratio of women's sections containing the word to the total number of women's sections; column C calculates the ratio of men's sections not containing the word to the total number men's sections. Column A sums columns B and C to produce the Zeta scores. Columns H-CA of row 1 show the number of different words (word types) in each section, and below them, the percentage of types that are marker words for women or men (these are not meaningful until the macro has operated). In cells F2-F3 the user can set the number of marker words for each group at different levels for sections of different sizes and can see three sets of results at once in H-CA.

Figure 1

An Excel macro automates the calculation of Zeta. Word frequency lists for the texts to be analyzed are entered into five sub-sheets (not shown). One contains the sections of the primary group and one contains the sections of the secondary group. Two more sub-sheets contain any independent sections by the primary and secondary groups (these can be used to test the method's success on known texts). Finally, there is a sub-sheet for the texts to be tested. The macro clears out old data, enters formulas into columns H-CA, rows 2-7, copies the texts out of the

sub-sheets into the main sheet, shrinks the columns, and enters ranks for the words in column F. It then sorts the words on their Zeta scores in column A, descending, so that the words most distinctively used by women appear at the top and those most distinctively used by men are at the bottom. It selects the 1000 most distinctive men's words and resorts them in reverse order, with the most distinctive at the top of their section. The sheet can handle 15,000 words, but the sample above analyzes 14,000 words (calculated in cell B2), so rows 11-1010 will contain the 1000 most distinctive women's words and rows 13,011-14,010 the 1000 most distinctive men's words.

Figure 2 shows the data after the macro runs. Here *mother's*, found in 13 of 26 sections by women and absent from 24 of 26 sections by men, is the most distinctive women's word in this comparison. The most distinctive men's word, *cross*, is found in just 3 of 26 sections by women, but is absent from only 10 of 26 sections by men. The most distinctive words for each group are found in columns CB and CC in descending order of distinctiveness. The figures in rows 2-7 of columns H-CA show how these distinctive words are distributed in each section. For example, H2-H3 shows that almost 15% of the words Derricote uses in this section are among the 500 most distinctive women's words, but only about 5% are among the 500 most distinctive men's words. For Ammons (1), in AH2-AH3, the first section by a man, the proportions are roughly reversed.

Figure 2

Columns BH-CA contain the same information for the test sections, 25 sections of poetry by seven female and seven male poets whose work had no part in the creation of the word list on which the analysis is based. Finally, the macro creates a scatter graph of the analysis with section names as labels (Fig. 3). The vertical axis records the proportion of types in each text that are marker words for men and the horizontal axis does the same for marker words for women. As Fig. 3 shows, twenty of the twenty-five new sections of poetry (80%) trend toward the group that matches their gender. The fact that these same words produce a similar, though slightly less accurate, result for seven male and seven female contemporary novelists is further evidence that the method is capturing some kind of genuine difference. This is a strong result, especially given the relatively small samples and the wide diversity that characterizes 20th century

American poetry. The little chart that follows Fig. 3 shows some of the clusters of related words that occur in the women's and men's marker words, and hints at the kinds of analysis that Craig's Zeta makes possible.

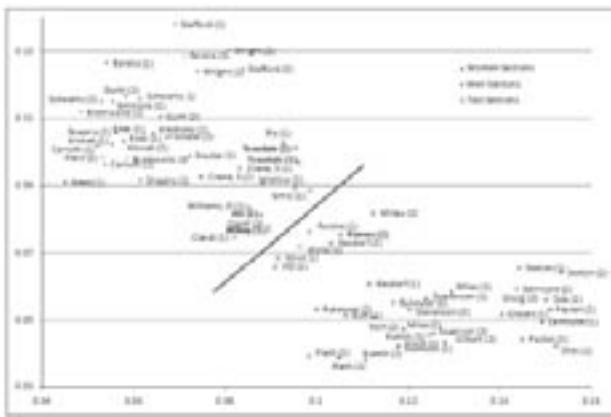


Figure 3

Cluster	500 Women's Markers	500 Men's Markers
Family	<i>mother's, father's, mother, father, children, ancestral, aunt, baby, birth, child, child's, cousins, daughters, family, generations, uncles</i>	
Religion	<i>altar, nuns, and praying</i>	<i>faith, heaven, hell, prayers, souls, spirit, Christ, gods, myth, paradise, religion, spirits, temple</i>
Houses/ Furniture	<i>danced</i>	<i>song, dancing, sing, dance, sang, dancer, music, singer, singing, sings</i>
Personal Pronouns	<i>he'll, I'd, mine, ourselves, she'd, she's, they'd, you'd, you're, yourself</i>	

References

- Burrows, J. F.** (2006). 'All the Way Through: Testing for Authorship in Different Frequency Strata'. *LLC*. **22**: 27-47.
- Burrows, J. F.** (2005). 'Who wrote Shamela? Verifying the Authorship of a Parodic Text'. *LLC*. **20**: 437-450.
- Craig, H., and Kinney, A., eds.** (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.
- Hoover, D.** (2007a). 'Corpus Stylistics, Stylometry, and the Styles of Henry James'. *Style*. **41**: 174-203.

Hoover, D. (2007b). 'Quantitative Analysis and Literary Studies'. *A Companion to Digital Literary Studies*. Susan Schreibman, Ray Siemens (eds.). Oxford: Blackwell, pp. 517-33.

Hoover, D. (2008). 'Searching for Style in Modern American Poetry'. *Directions in Empirical Literary Studies: Essays in Honor of Willie van Peer*. Sonia Zyngier, et. al. (eds.). Amsterdam: John Benjamins, pp. 211-27.

Hoover, D. (2009). 'The Excel Text-Analysis Pages'. <http://https://files.nyu.edu/dh3/public/The%20Excel%20Text-Analysis%20Pages.html>.

Hoover, D. (2010: forthcoming). 'Authorial Style'.

The Dickens Lexicon and its Practical Use for Linguistic Research

Hori, Masahiro

hori@kumagaku.ac.jp

Kumamoto Gakuen University

Imahayashi, Osamu

imahaya@hiroshima-u.ac.jp

Hiroshima University

Tabata, Tomozi

tabata@lang.osaka-u.ac.jp

Osaka University

Nishio, Miyuki

nishio@hiro.kindai.ac.jp

Kinki University

It was not until the beginning of World War II that Dr. Tadao Yamamoto first established a plan for the compilation of the *Dickens Lexicon* in his mind; the earliest plan of which was suggested in *Studies in English Literature* (Vol. XIII, No. 3, 1943). As the war situation turned progressively worse, the completion of the Lexicon was left to future efforts. He decided, however, to make an "Introduction" to it in the early spring of 1944, and in the same year presented it as a doctoral thesis to the University of Tokyo under the title of *Growth and System of the Language of Dickens: An Introduction to A Dickens Lexicon*, for which he obtained the degree of Doctor of Literature from the University in 1946. The dissertation was first published in 1950 by Kansai University Press through the generous efforts of the late Professor Jiichi Hattori at Kansai University, and with financial support from the English Philological Society of Kansai University. In 1953 he was awarded the Japan Academy Prize for this book. The second edition and "An index to Tadao Yamamoto's *Growth and system of the Language of Dickens: With supplementary notes & corrections*" were published separately by the same press in 1952. The third revised edition was published by Keisisha Publishing Company in 2003.

In 1948 Dr. Yamamoto organised the first joint research for the compilation of *A Dickens Lexicon*, which was granted a Government Subsidy for Scientific Research by the Department of Education for 1948. The members of the joint research mainly consisted of his pupils in Hiroshima University of Literature and Science. The members chose one of Dickens' works and collected the materials for the *Lexicon*. The participants and their selected works are as follows:

Tadao Yamamoto	<i>Oliver Twist</i>
Michio Masui	<i>Bleak House</i>
Chiaki Higashida	<i>A Tale of Two Cities</i>
Tamotsu Kurose	<i>Christmas Books</i>
Hiroshige Yoshida	<i>Nicholas Nickleby</i>
Masami Tanabe	<i>Old Curiosity Shop</i>

After he moved to Osaka Women's University in 1952, Yamamoto organised the second joint research for the compilation of the *Dickens Lexicon*; the members of which included Michio Masui, Chiaki Higashida, Tamotsu Kurose, Haruo Kouzu, Yasuo Yoshida, Tadahisa Goto, Jun Matsumoto, Tamotsu Matsunami, Hideo Hirooka, and Michio Kawai. The joint research was granted a Subsidy for Government Scientific Research by the Department of Education for 1952. The process and result of it were reported in *Anglica* (1954: 438-9) as follows:

As a preliminary work for the compilation of the *Dickens Lexicon* we aimed at establishing the working principles of selecting materials for our research. For this purpose each of the members chose one of Dickens' writings from which necessary materials should be extracted. It was desired that each participant should at the outset prepare explanatory notes to the work chosen and as the next step offer slips of quotations under separate items with comments if necessary.

<i>Sketches by Boz</i> (B)	Matsunami
<i>Pickwick Papers</i> (P)	Matsumoto
<i>Christmas Carol</i> (Carol)	Kurose
<i>Martin Chuzzlewit</i> (MC)	Masui
<i>Cricket on the Hearth</i> (Cricket)	Goto
<i>Dombey and Son</i> (DS)	Higashida
<i>David Copperfield</i> (DC)	Yoshida
<i>A Tale of Two Cities</i> (TC)	Imagawa
<i>Great Expectations</i> (GE)	Ishino

Separately the present writer has prepared a collection of detailed notes to *Oliver Twist* (OT), with which materials chosen out of the above works are to be collated.

Slips collected amount to 6504, from which 2915 have been sifted and adopted for the present research. They may be roughly classified as follows:

01.	Names and subjects	472
02.	Word-forms	152
03.	Slang and dialects	388
04.	Quotations and allusions	256
05.	Expressions coming from some definite situations or surroundings	337
06.	Phrasal expressions	200
07.	Exclamations, asseverations, swearing, &c.	174
08.	Intensive expressions	88
09.	Precise and energetic expressions	73
10.	Those with bodily names	31
11.	Miscellaneous	519
12.	Words and phrases particularly collated with the notes to Oliver Twist	225
	Sum total	2915

In Yamamoto's conclusion, he commented on the limitations and difficulties of this joint research as follows:

"... as a joint work ours for this time has remained at the very tentative stage. It has taught us that the desideratum is a perfect team-work with sufficient preparation and training that cost us an enormous amount of time and labour. With all our efforts, however, we must admit that we continually suffer from the considerable limitation of our knowledge, and under the present conditions there are insurmountable difficulties in having access to each and every requisite source of information. It would indeed be a consummation devoutly to be wished if we could come directly in touch, not exclusively through the narrow channel of written sources now at our disposal, with all things that have conspired to create Dickens and his language." (451)

The research team was, however, broken up, and a new downsized one was organised. Its members were Chiaki Higashida, Yasuo Yoshida, Jun Matsumoto, and Shigekiyo Kawahara. The result was published in *Dickens no Buntai* (*Dickens' Style* in English) from Nan'un-do in 1960, but this joint research did not bear fruit either. From that time Yamamoto began to collect the materials for the *Dickens Lexicon* once again from *Pickwick Papers* all by himself, but unfortunately on the 28th of July in 1991, he died without seeing it accomplished.

This poster session is an interim report on the *Dickens Lexicon* project, which was newly organized in 1998 by a research group of twenty scholars whose ultimate aim has been to compile the *Dickens Lexicon* from approximately 60,000 cards, which Dr. Tadao Yamamoto (1904-91) elaborately drew up and left to us. The *Dickens Lexicon* is expected to be released as the "Dickens Lexicon Online" on an Internet website with a multifunctional search engine, in the near future. This poster session provides an introduction

to the *Dickens Lexicon* project, including its practical use for research.

The *Dickens Lexicon* is designed as a web-based reference resource. Users will be able to search and retrieve lexical data (an idiom, its word class, definition, source, and quotation), stored in the original card database of approximately 60,000 indexed entries without installing extra software (apart from a web browser) on their computers. The lexicon will also be implemented with a multifunctional information retrieval system. In addition to the indexed entries, the lexicon will make it possible to retrieve frequency information on lexical items (from single words to phrases, including multi-word units) drawing upon the full corpus of Dickens' texts and an additional set of major 18th and 19th century fictional texts. A range of functions such as concordance display, sort capability, and distribution chart will be available in a user-friendly interface. Therefore, a close scrutiny of idioms appearing in the *Dickens Lexicon* with a multifunctional information retrieval system will not only make us aware of the ways idioms provided an important characteristic in Dickens' usage of English, compared with those in other major 18th and 19th century fictional texts, but will also provide insights into the characteristic structure of idiomaticity in the English language as well.

References

Yamamoto, Tadao (1950 [2003]). *Growth and System of the Language of Dickens: An Introduction to A Dickens Lexicon*. Hiroshima, Japan: Keisuishisha.

Dingler-Online – The Digitized "Polytechnisches Journal" on Goobi Digitization Suite

Hug, Marius

marius.hug@culture.hu-berlin.de
Humboldt-Universität zu Berlin

Kassung, Christian

CKassung@culture.hu-berlin.de
Humboldt-Universität zu Berlin

Meyer, Sebastian

sebastian.meyer@slub-dresden.de
SLUB-Dresden

This project located at Humboldt-Universität zu Berlin sets out to digitize Dingler's "Polytechnical Journal" ("Polytechnisches Journal"), 1820-1931. Aside from the digitization of the journal's images, we encode the OCRed text according to the Text Encoding Initiative Guidelines TEI-P5. Our online edition of Dingler's journal will be freely available on a state-of-the-art system called Goobi Digitization Suite, which is a new production and presentation solution funded by the DFG (German Research Foundation).

1. Dingler's Journal

In 1820 the German chemist and industrialist J. G. Dingler started publishing the "Polytechnisches Journal". This journal was to include a personal but representative selection of a broad variety of articles. Originally most of these articles had been published in magazines all over Europe and, though most of them originated from the UK, there were also specimens from France, Italy, and Russia.

The "Polytechnisches Journal" was published over a period of 111 years. Thus, the journal became an extremely important source for the history of 19th century knowledge, as it is an account of period that included the industrialization, the progress of transport and communication, and the differentiation of various technologies. For instance, the journal covers the discovery of electromagnetism by Hans Christian Oersted (1820) and the theory of relativity by Albert Einstein (1905/1915). It contains articles on steam engines and locomotives, as well as bicycles and automobiles.

Synchronic and diachronic transfer of knowledge and technique

Articles published in Dingler's journal give us an example of the emergence of culture in a technical context. In the process of industrialization, new technical achievements profoundly affected everyday life. It is in the interplay of science and knowledge that the journal evolves its epistemic significance. The "Polytechnisches Journal" is unique and highly relevant for very different research fields which focus on the cultural history as it emerged from Europe's technical transformations. It is significant not only for people engaged in the history of science but for anyone interested in the cultural heritage of Europe.

2. Dingler-Online – The encoding

Linking

Since Dingler annotated his editorial work very thoroughly, we find all necessary metadata on each article contained within the journal. He even went one step further: Dingler cross-referenced other source material on issues inherent to each article. Therefore "linking" is one of the main tasks for enriching the text. Doing this consistently from the very beginning of our project we are aiming at a network of digitized knowledge for that period covering the whole of Europe. Any digitized magazine of a somehow technical background of the 19th century will be interesting to be linked to.

Indexing

As is true for any non-digitally published magazine, researching the contents of journals is a time consuming task. Right from the beginning Dingler knew he would have to give assistance to those accessing his material, so he compiled an index once a year. In 1843 the first so-called "Real-Index" was published, a third-hand work which covered the first 78 volumes of the journal. All in all there are four of these "Real-Indexes".

Based on these two different kinds of indexes as well as our index-related TEI-encoding, we will be able to provide a deeply granulated and dynamically generated index. It will consist of a register of persons (differentiated according to their role, i.e. author, translator, originator etc.), objects, and, among others, those journals, which were the source of the published articles.

The articles – our key component

Dingler's journal comes in 360 volumes, each including 4 to 6 issues. The key components of our edition are the 50 to 170 articles in each volume. Even at this very basic level, we distinguish between two types of articles, since there are in fact a couple of articles published for the first time in the "Polytechnisches Journal" in addition to the reprinted articles. We extract all these articles from the volume and provide access to downloadable PDF-versions as well as different formats of established

bibliographical meta-data. In the long-run we aim at providing access to PDFs generated dynamically via XSL-FO.

Text and images

The editors of the journal strictly adhered to the medial conditions of their time. In 1820 Dingler started using Gothic typescript for text, and copper engravings for the imprints. In later issues (starting in the 1870s) we find Antiqua letters and floating images integrated within the text.

Our aim is a re-interpretation of the relationship between text and images. Dingler completed each volume with technical drawings and visualizations on additional plates. Hence, up to 40 figures on a plate are encoded according to their specific coordinates using the Image Markup Tool developed by the University of Victoria. Via hyperlink we are able to provide access to a zoomable view of each figure. This approach has two immediate advantages: Firstly, it enables parallel reading of text and image and therefore adopts the original layout, in which plates were attached to the back of each volume as foldouts. Secondly, we can provide a new kind of readability. For economic reasons the plates were densely packed with images. Highlighting them per mouseover will be much more convenient, allowing to inspect them in more detail and thus enabling a wider integration of the text and images.

Since right from the beginning Dingler insisted on very detailed and thus expensive lithographs rather than wood engravings, we have made it our task not to veer from the standard set by Dingler at this point.

3. Dingler-Online II – The appearance

Not only since we are facing the challenging task of digitizing Gothic type in more than 220 volumes of the "Polytechnisches Journal", we find ourselves in good company with two rather impressive German digitization projects: Grimms "Deutsches Wörterbuch" and Krünitz's "Oeconomische Encyclopädie". Both made use of double-keying and therefore sent their books/images to Asia where the text digitization took place. Afterwards so-called TUSTEP-routines were employed in order to match the two different text versions.

In the following we will take a closer look at different aspects, which will take our project one step further than the aforementioned approaches.

3.1. Goobi Digitization Suite

With the so called Goobi Digitization Suite – a software solution funded by the DFG and developed

by the SLUB-Dresden (Sächsische Landesbibliothek – Staats- und Universitätsbibliothek) and the SUB-Göttingen (Niedersächsische Staats- und Universitätsbibliothek) – we will be using a completely new technology on the market.

The Goobi Suite consists of two parts: *Goobi.Production* and *Goobi.Presentation*. *Goobi.Production* is a web-based tool for managing a digitization workflow using Java technology. Among other features, it comes with a very flexible metadata editor, an user-based permission system, and visually enhanced statistics.

Since at the beginning of our project the Goobi Suite wasn't available yet, we found an experienced service provider for text digitization and (semi-)automatic encoding: the Editura GmbH. Their OCR produces very good results even for Gothic type, given that the images are scanned at 600 dpi.

Editura encodes the OCRed text and already enriches it according to the TEI-P5 guidelines. This step includes 'tagging' the structure and special attributes of the text to an encoding level between 3 and 4. Thus the digitization of the text includes more than a basic structural encoding and we can concentrate on a more scientific encoding approach going beyond other projects comparable in extent.

Apart from XML-files in TEI-encoding and images encoded using the Image Markup Tool our service provider delivers elaborate METS-files which are necessary for a presentation of the edition in the so called DFG-viewer, as well as in *Goobi.Presentation*, which we use as part two of the Goobi Digitization Suite. This is a full-featured web presentation layer for digital material and is based on the TYPO3 CMS Framework, which can hold a regular website, too. Hence, *Goobi.Presentation* integrates perfectly into any page inside the CMS.

The whole software suite is considered open source and freely available to everyone. As can be seen in our project, *Goobi.Presentation* can be used independently from *Goobi.Production*. This modularity of Goobi is ensured by the consequent usage of the international standards METS, MODS and TEI.

3.2. Customizing Goobi

The more data there is to present, the less important any unstructured information becomes. This is why encoding and a directed access to data, via searching or browsing, becomes more and more important.

Goobi.Presentation makes it possible to customize the search engine. Naturally one will be able to search any term anywhere in the text. In addition, it is possible to limit the search results referring to

different issues. For instance: if someone is looking for all articles on patent applications on steam engines published in the magazine in the 1840s, they will just have to search for "steam engine", then restrict their search to "text type" patent application, and "time" 1840s.

4. Conclusion

Dingler-Online is an enriched digitization that is neither simply image-based nor massproduced. It is a user-friendly platform which inspires a broad use not restricted to historians of technology or, come to that, researchers, but is open to the interested public in general.

References

- Das Deutsche Wörterbuch von Jacob und Wilhelm Grimm auf CD-ROM und im Internet.* <http://germazope.uni-trier.de/Projects/DWB/>.
- Dingler-Online | Das digitalisierte Polytechnische Journal.* <http://www.polytechnischesjournal.de/>.
- Editura GmbH.* <http://www.editura.de/>.
- Fischer, F.** (2007). 'Dinglers Polytechnisches Journal bis zum Tode seines Begründers (1820-1855)'. *Archiv für Geschichte des Buchwesens*. **15**: 1027-1142.
- Goobi – DigitalLibraryModules.* <http://www.goobi.org>.
- Oeconomische Encyclopädie online.* <http://www.kruenitzl.uni-trier.de/>.

The MLCD Overlap Corpus (MOC)

Huitfeldt, Claus

Claus.Huitfeldt@uib.no

Department of Philosophy, University of Bergen

Sperberg-McQueen, C. M.

cmsgmcq@blackmesatech.com

Black Mesa Technologies LLC, USA

Marcoux, Yves

ymarcoux@gmail.com

Université de Montréal, Canada

For some time, theorists and practitioners of descriptive markup have been aware that the strict hierarchical organization of elements provided by SGML and XML represents a potentially problematic abstraction. The nesting structures of SGML and XML capture an important property of real texts and represent a successful combination of expressive power and tractability. But not all textual phenomena appear in properly nested form, and for more than twenty years students of markup have been exploring methods of recording overlapping (non-hierarchical) structures. Useful surveys include (Barnard et al. 1995), (DeRose 2004), and (Witt et al. 2005).

Some approaches to the overlap problem take the form of non-SGML, non-XML syntaxes and non-tree-like data structures. One example is offered by the TexMecs syntax and Goddag data structures proposed by the project Markup Languages for Complex Documents (MLCD) based at the University of Bergen. Another is the Layered Markup and Annotation Language (LMNL). A third is the so-called multi-colored trees defined by (Jagadish et al. 2004).

Other approaches exploit the optional *concurrent markup* feature of SGML (Sperberg-McQueen and Huitfeldt 1998), or apply it, with suitable modifications, to XML (Hilbert et al. 2005).

But by far the largest number of published approaches to problems of overlapping markup involve the use of SGML and XML themselves to record the information. They exploit the semantic openness of SGML and XML to supply non-hierarchical interpretation of what are often thought to be inescapably hierarchical notations.

The SGML/XML-based approaches to overlap fall, roughly, into three groups: milestones, fragmentation-and-reconstitution, and stand-off annotation. Milestones (described as early as

(Barnard et al. 1988), and used in (Sperberg-McQueen and Burnard 1990) and later versions of the TEI *Guidelines*) use empty elements to mark the boundaries of regions which cannot be marked simply as elements because they overlap the boundaries of other elements. More recently, approaches to milestone markup have been generalized in the Trojan Horse and CLIX markup idioms (DeRose 2004).

Fragmentation is the technique of dividing a logical unit which overlaps other units into several smaller units, which do not; the consuming application can then re-aggregate the fragments.

Stand-off annotation addresses the overlap problem by removing the markup from the main data stream of the document, at the same time adding pointers back into the base data. Many language corpora use forms of stand-off markup (e.g. (Carletta et al. 2005), (Witt et al. 2005), (Stührenberg and Goecke 2008)).

For all the variety of methods and proposals for handling overlap, there is remarkably little consensus on the best approach. Even systematic comparisons are scarce, although several of the surveys provide at least a broad categorization of methods. Partly this reflects a pragmatic issue (many methods used in production systems are devised for use by specific projects, which do not wish to engage in a systematic comparison of interest to markup theorists, but to get on with their discipline-specific work); partly it reflects a difficulty in comparing different schemes point to point, owing to the scattered and informal nature of the documentation.

And finally, despite the work of the last twenty years we still have only an incomplete understanding of the different structural and semantic forms of overlapping structure, and the implications for markup practice of different forms of overlap. The pervasive but unsystematic overlap of verse and dramatic structure in verse drama, or of formal and physical structure in any printed book, seems to present one kind of phenomenon. The occasional but richly significant overlap of structures characteristic of enjambement in verse may appear, on the other hand, to be of a different kind. Is it?

The MLCD Overlap Corpus (MOC) is intended to make it easier to compare different methods of handling overlap, not just on theoretical or abstract grounds, but in terms of concrete examples from real and constructed texts. The essential idea of the corpus is to make available a single body of material, ranging from compact examples to full texts of novel or five-act-play length, tagged for the same information, using a variety of overlap notations.

Consider the following simple example (from (Hilbert et al. 2005)) of a discourse situation in which

the utterance structure overlaps with the syntactic structure:

Peter: Hey, Paul! Would you give me

Paul: the hammer?

(Hilbert et al. 2005) give the following representation of this example in the notation now known as XConcur (then MLX).

```
<!DOCTYPE (1)div SYSTEM "tei/dtd/teispok2.dtd">
<!DOCTYPE (2)text SYSTEM "tei/dtd/teiana2.dtd">
  <(1)div type="dialog" org="uniform">
    <(2)text>
      <(1)u who="Peter">
        <(2)s>Hey Paul!</(2)s>
        <(2)s>Would you give me
      </>
      <(1)u who="Paul">
        the hammer?</(2)s>
      </>
    </>
  </>
</>
```

Using the CONCUR feature of SGML, a very similar representation can be given (elided here for space reasons). It might be represented in TexMecs this way:

```
<div type="dialog" org="uniform">
  <u who="Peter">
    <s>Hey Paul!</s>
    <s SID="s2"/>Would you give me
  </u>
  <u who="Paul">
    the hammer?<s eID="s2">
  </u>
</div>
```

The goal of MOC is to make examples available in a broad variety of notations, as well as those just given:

- various forms of TEI markup, using different TEI mechanisms (*next* and *prev* attributes, the *part* attribute, virtual elements, stand-off markup using feature structures, etc.)
- TexMecs (Huitfeldt and Sperberg-McQueen 2003)
- XStandoff (Stührenberg and Jettka 2009)
- Multix (Chatti et al., 2007)
- Sekimo General Format (SGF) (Stührenberg and Goecke 2008)
- Nite (Carletta et al. 2005)
- Earmark (Di Iorio et al. 2009)

There will be three sets of data:

- twenty or more 'toy' examples like the one just given, typically just a few lines in length. Most of the toy examples will be drawn from existing literature on overlap; almost all of them will be

constructed texts, though some will be very short extracts from literary or other natural texts.

- ten or more 'short' examples, typically corresponding to a few pages of printed material, mostly extracts from natural texts.
- five or more 'long' examples, full-length natural texts. We will draw these partly from an existing collection of literary texts used as a test bed for full-text software and partly from existing language corpora.

The toy examples will be tagged manually in the various notations selected. The short examples will be tagged using semi-automated processes (i.e. partly by hand and partly automatically), and checked carefully for correctness. The long examples will be tagged using mostly automated processes, and checked carefully for correctness.

Since the purpose of MOC is to illuminate problems connected with overlap and with existing proposals for handling it, there will be no attempt to make the selection of texts representative of any particular natural language community. The relevant population is not a particular set of natural-language users, but the set of people who work with natural-language texts for various purposes. In such a small corpus, we cannot and do not hope for statistical representativeness, but only for an illuminating variety of examples. Accordingly, we will seek to include examples illustrative of problems encountered in:

- literary and lexicological study
- metrical study
- language corpora (discourse analysis, syntax, prosody, ...)
- change markup and multi-versioned texts
- historical-critical editions
- analytic bibliography
- historical annotation

Apart from simply illustrating the ways in which different notations represent the same information, MOC should provide sample test data useful for a variety of tasks and studies:

- development of automatic translation among notations (the existing samples of the target notation serve as comparison points for the results achieved by the automatic translator)
- development of software intended to handle any of the notations represented
- construction of domain-appropriate queries against the various notations (does notation N1

make it easier to construct suitable queries than notation N2?)

- comparative measures of markup complexity
- analysis of different kinds and forms of overlap: do structural patterns vary with different kinds of markup? Do the domain-specific implications of overlap (and thus the domain-oriented requirements for manipulating the data) vary?
- development of tools for automatic extraction of formalized representations of the meaning of markup

Performance comparisons are notably missing from this list; MOC will be too small to provide performance measurements relevant to searches across typical modern collections in the gigabyte size range. (On the other hand, the long samples may be useful for at least preliminary performance comparisons and preparation for more large-scale testing.)

At the time this abstract is prepared, the first version of MOC is expected to be partially completed before the DH 2010 conference; the presentation will include an account of the work to date, problems encountered, and a forecast of the work remaining before completion of the corpus.

Follow-on work includes experimentation with existing full-text indexing and query systems to test the different characteristics of different markup styles on query formulation and retrieval time; we also expect to work on automated translations among various notations.

References

- Barnard, D., Hayter, R., Karababa, M., Logan, G. and McFadden, J.** (1988). 'SGML Markup for Literary Texts'. *SGML Markup for Literary Texts*. **22**: 265-276.
- Barnard, D., Burnard, L., Gaspart, J. P., Price, L. A., Sperberg-McQueen, C. M. and Varile, G. B.** (1995). 'Hierarchical encoding of text: Technical problems and SGML solutions'. *Computers and the Humanities*. **29**: 211-231.
- Carletta, J., Evert, S., Heid, U. and Kilgour, J.** (2005). 'The NITE XML Toolkit: data model and query'. *Language Resources and Evaluation*. **39(4)**: 313-334. <http://doi:10.1007/s10579-006-9001-9>.
- Chatti, N., Kaouk, S., Calabretto, S. and Pinon, J. M.** (2007). 'MultiX: an XML-based formalism to encode multi-structured documents'. *Proceedings of Extreme Markup Languages 2007*. Montréal (Canada), Aug.

2007. <http://conferences.idealliance.org/extreme/html/2007/Chatti01/EML2007Chatti01.html>.
- DeRose, S. J.** (2004). 'Markup overlap: A review and a horse'. *Proceedings of Extreme Markup Languages 2004*. Montréal (Canada), Aug. 2004. <http://conferences.idealliance.org/extreme/html/2004/DeRose01/EML2004DeRose01.html>
- Di Iorio, A., Peroni, S. and Vitali, F.** (2009). 'Towards markup support for full GODDAGs and beyond: the EARMARK approach'. *Proceedings of Balisage: The Markup Conference 2009*. Montréal (Canada), August 11-14, 2009. <http://www.balisage.net/Proceedings/vol3/html/Peroni01/BalisageVol3-Peroni01.html> <http://doi:10.4242/BalisageVol1.Stuehrenberg01>.
- Hilbert, M., Schonefeld, O. and Witt, A.** (2005). 'Making CONCUR work'. *Proceedings of Extreme Markup Languages 2005*. <http://conferences.idealliance.org/extreme/html/2005/Witt01/EML2005Witt01.xml>.
- Huitfeldt, C. and Sperberg-McQueen, C. M.** (2003). *TexMECS: An experimental markup meta-language for complex documents*. University of Bergen. <http://decentius.aksis.uib.no/mlcd/2003/Papers/texmecs.html>.
- Jagadish, H.V., Lakshmanan, L. V. S., Scannapieco, M., Srivastava, D. and Wiwatwattana, N.** (2004). 'Colorful XML: one hierarchy isn't enough'. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. Paris (France), pp. 251-262. <http://doi.acm.org/10.1145/1007568.1007598>.
- Sperberg-McQueen, C.M. and Huitfeldt, C.** (1998). 'Concurrent Document Hierarchies in MECS and SGML'. *Literary and Linguistic Computing*. **14**: 29-42.
- Stührenberg, M. and Jettka, D.** (2009). 'A toolkit for multi-dimensional markup: The development of SGF to XStandoff'. *Proceedings of Balisage: The Markup Conference 2009*. Montréal (Canada), August 11-14, 2009. <http://www.balisage.net/Proceedings/vol3/html/Stuhrenberg01/BalisageVol3-Stuhrenberg01.html> <http://doi:10.4242/BalisageVol1-Stuehrenberg01>.
- Stührenberg, M. and Goecke, D.** (2008). 'SGF — An integrated model for multiple annotations and its application in a linguistic domain'. *Proceedings of Balisage: The Markup Conference 2008*. Montréal (Canada), August 12-15, 2008. <http://www.balisage.net/Proceedings/vol1/html/Stuehrenberg01/BalisageVol1-Stuehrenberg01.html> <http://doi:10.4242/BalisageV01.Stuehrenberg01>.

Creative Engagement with Creative Works: a New Paradigm for Collaboration

Jones, Steven E.

sjones1@luc.edu

Loyola University Center for Textual Studies and Digital Humanities

Shillingsburg, Peter

Loyola University Center for Textual Studies and Digital Humanities

Thiruvathukal, George K.

Loyola University Center for Textual Studies and Digital Humanities

Funded in part by a Digital Start-up Grant from the National Endowment for the Humanities, **Creative Engagement with Creative Works** is a project to build a new online environment (e-Carrel) and integrated tools with the aim of improving understanding of creative processes across various humanities disciplines and genres. Studies of the production and reception of literary, historical, musical, and philosophical works are all built on primary materials that are textual in the broad sense--documentary, material objects: manuscripts, newspapers, periodicals, pamphlets, books, and images. But current solutions to digitizing and providing access to these materials are structurally flawed and lead away from the often-stated goal of extensive collaboration. Digital transcriptions relying on XML or other inline markup can often prevent or limit collaboration on the files themselves, can (paradoxically) threaten a project's integrity, and can lead to early maintenance problems and the ultimate abandonment of projects beyond the lifetimes of their initial creators.

We propose using standoff markup instead, indexing these files to inviolate core files, thus affecting the way humanities disciplines interact with primary materials--moving away from proprietary, look-but-don't-touch window-case projects toward secure and enduring projects that are open to ongoing annotation and re-markup, and which thus encourage widely collaborative knowledge sites.

Electronic projects that restrict their construction and enhancement to a select few persons as a way to protect the intellectual integrity of their resources inadvertently introduce a significant threat to the durability of text files, whose hard-won accuracy is the result of expert attention. The creators' anxiety to protect the results of their time-consuming and

labor-intensive in-line-tagging leads to a proprietary attitude that's detrimental to scholarly collaboration, because everyone knows that the text files are vulnerable to inadvertent change every time they are reopened for further tagging. Protective restriction also restricts truly collaborative work, the life of digital humanities projects in the world.

These restrictions can be loosened or eliminated by a fundamentally different approach to collaboration, durability, and maintenance, pioneered in scientific fields following principles of:

- modular component structure,
- connectivity,
- extensibility,
- distribution and aggregation systems,
- stand-off enhancement mechanisms, and
- methods for identifying and crediting researchers with their individual contributions to composite research projects.

These trends in software development allow for decentralized alterations. But even free/open source (FOSS) software projects are not without problems. The past was dominated largely by projects where the code was kept under tight wraps, using version control systems such as CVS (Concurrent Versions System) and Subversion. Only a handful of FOSS projects succeed with this model, notably the Firefox project, but many FOSS projects are already migrating to more distributed approaches (the Python language, the Linux kernel, etc.). Much like archives, software development projects in the FOSS community tend to self-organize and establish their own governance. While they make their code available (as required by the FOSS licensing schemes) they also tend to fall into disuse, making it hard for new developers to come along and take the project to new levels.

Distributed version control systems record a project's entire history and state, which can be copied freely, allowing derived works to take place. When a copy is made, however, the entire history is kept intact, allowing new contributors to either make their own changes or to push their changes back to the original maintainer. The push/pull model is so sophisticated that anyone who makes changes can get recognition for their work, because their specific changes to the code are encoded in the derived history.

Our Creative Engagement with Creative Works and e-Carrel environment incorporates these principles and adds the functions of stand-off files for markup and annotation, along with a dynamic authentication mechanism. Experimentation has shown the considerable promise of such functionality. (See the Just InTime Markup system prototyped at

the University of New South Wales at ADFA by our Senior Consultant, Paul Eggert and his team working on Australian literature; and see Desmond Schmidt's MVD or Multi Version Document system, the architecture and some code of which we are incorporating). These are the building blocks of our project's coherent vision for archiving creative works for creative collaboration, preservation, and dynamic interactive access, realized in the form of tools and programming frameworks.

We establish an image file and a base text for each significant version of a work. Text data for all versions are compressed in a composite, inviolable CorTex file, which anchors all stand-off contributions. Each participant's contribution is credited and protected from work by other contributors. Endusers choose a historical text (or other object) plus desired types of enhancements from a menu dynamically aggregated from distributed sources. The e-Carrel processes and presents perspectives of texts and enhancements for viewing, printing, or export in commonly used formats. Individual projects using the system can vet and certify parts of a project. And--most important--this system allows for the storage, retrieval, and coordination of different, even conflicting editorial or critical approaches to the same literary work. In this way, it opens the horizon of any given project's ongoing reception.

By giving scholars and students a vested interest in a growing integrated collaborative project, CECW ensures preservation and access to textual research projects and their superstructure of critical analysis at the same time that it promotes collaboration beyond the project initiators' participation and goals. Project viability follows community ownership and becomes a widely distributed responsibility. The system ensures long-term maintenance and growth through collective ownership, distributed storage, and the principle of LOCKSS (Lots of copies keep stuff safe). Our strong definition of LOCKSS, which distributes multiple accurate copies, and in an ongoing process verifies them by way of the persistent CorText data and checksum system, is significantly different from the "soft" version of LOCKSS that just sends "copies" into the world in whatever state and trusts to the hive mind for endurance and integrity.

CECW makes use of RDF capabilities and is XML aware but is by design markup agnostic. It allows for the importation of materials from other systems and prepares perspectives of the E-Carrel materials for export in various forms compatible with other systems (such as PDF and in-line coded XML).

This poster session will provide demos of portions of the software in development, in particular, the standoff markup tool used within the E-Carrel

environment. A case-study CoreText will be loaded for the purposes of demonstration.

References

- Eggert, Paul** (2009). 'The Book, the e-Text, and the Worksite'. *Text Editing, Print and the Digital World*. Ashgate.
- Loyola University Center for Textual Studies and Digital Humanities.* <http://www.ctsdh.luc.edu/>.
- Schmidt, Desmond** (2009). 'A Data Structure for Representing Multi-Version Texts Online'. *Int. J. Human-Computer Studies*. **67**: 497-514.
- Shillingsburg, Peter** (2006). *From Gutenberg to Google*. Cambridge: Cambridge UP.

Distant Reading and Mapping Genre Space via Conjecture-based Distance Measures

Juola, Patrick

juola@mathcs.duq.edu
Duquesne University

One of the key problems facing digital humanities today is the increasing number and size of digital repositories and the relative lack of tools for studying them. A collection of a million books (Crane, 2006) is no more useful than a collection of ten thousand if you can't read more than a hundred of them in a realistic timeframe. Scholars like Moretti (2005) have proposed a new analysis method, termed "distant reading," to enable computer-aided large-scale analysis of such collections. In previous work (Juola and Bernola, 2009), we have proposed using a conjecture generator (Conjecturator, see also <http://www.twitter.com/conjecturator>) as another computer-aided analysis method.

Underlying the Conjecturator is the idea that the computer can be deployed to autonomously generate "facts" about a given text repository. Like its predecessor and inspiration *Graffiti* (Fajtlowicz, 1988), the conjecturator generates template-based "conjectures" that might or might not be true about the repository and the texts in it. A sample conjecture might be something like:

- The concept of "archivist" appears more in mid-Victorian novels than in psychological realism novels or, more obviously,
- The concept of "femininity" appears more in feminist novels than in novels with gothic elements.

(Who would have thought, eh?)

As discussed in (Juola and Bernola, 2009), these simple conjectures can be easily and quickly tested to refute or confirm their validity. This enables the computer to quickly generate a pile of isolated "facts" about the text repository, but does not provide a useful framework for interpretation, explanation, or understanding (which still requires human expertise).

However, this "pile of facts" can provide useful source material for distant reading. In this paper, we demonstrate one way to extend this conjecture-based analysis to a large-scale "distant reading" and visualization of genre differences. Repeated

generation of conjectures will create a large catalogue of potential differences between any particular category pair, some true/supported, and some false. The number of "true" differences, or alternatively, the percentage of true differences, can be viewed as a distance between the categories, a distance measuring the degree of difference between the concepts commonly written about in those genres. If, for instance, "epistolary novels" differ in 26 significant ways from "tragic novels", but only in one significant way from "fiction of manners," we can consider "epistolary novels" to be a closer genre in terms of expressed concepts to "fiction of manners" than to "tragic novels." This high-order analysis gives us a large-scale conceptual grouping of genre categories.

To aid in the study of such differences, we compile the differences into a matrix and apply multidimensional scaling (MDS) (Cox and Cox, 2001). This statistical technique takes a high-dimensional data set defined by interpoint distances and embeds/rescales it to fit a smaller number of dimensions (in this case, two) while minimizing distortion. The resulting two-dimensional coordinates can be plotted to give a visual "map" of the space of genres. We demonstrate this technique using an enlarged set of 136 novels representing 36 genres (including time period and authorial attributes as "genres") and approximately 10,000 validated conjectures (culled from approximately 85,000 conjectures in total).

The resulting images clearly indicate that this method is a new and viable way of performing large-scale distant reading. As can be seen in Figure 1, the resulting "map" passes many obvious tests for rationality; for example, mid-Victorian novels represent an intermediate stage between early and late Victorian novels; similarly, "male authored novels" in general are an intermediate between "American male authored novels" and "English male authored novels," reflected exactly what intuition suggests. We leave it to genre specialists to examine the map in detail and to see whether actual genres equally reflect our intuitions. It is easy and relatively efficient to apply and almost entirely document-agnostic; it can be applied as easily to journal articles (and map the space of scholarship) or to newspaper corpora (perhaps mapping the space of editorial policies and politics) as to novel genres.

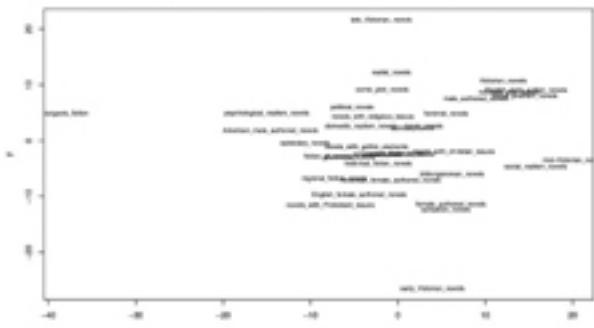


Figure 1

References

- Cox, T.F., Cox, M.A.A.** (2001). *Multidimensional Scaling*. Chapman and Hall.
- Crane, Gregory** (2006). 'What Do You Do With a Million Books?'. *D-Lib Magazine*. **12**(3).
- Fajtlowicz, Siemion** (1988). 'On conjectures of Graffiti'. *Discrete Mathematics*. **72**.
- Juola, Patrick** (2009). 'Mapping Genre Space via Random Conjectures'. Presented at DHCS-2009, IIT. Chicago, IL.
- Juola, Patrick, Bernola, Ashley** (2009). 'Conjecture Generation in the Digital Humanities'. *Proc. DH-2009*. 2009.
- Moretti, Franco** (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Psycholinguistically Plausible Events and Authorship Attribution

Juola, Patrick

juola@mathcs.duq.edu

Duquesne University

Authorship attribution (Juola, 2008) is an important emerging subdiscipline of digital scholarship, but it suffers from a lack of connection to other areas and disciplines, which in turn strongly limits both applicability and uptake. It is now unquestionable that computers can infer authorship attributes with high accuracy, but the accurate inference processes tend not to inform us about the actual authors (Craig, 1999). Among the best methods, for example, are the analysis of the most frequent function words such as prepositions (e.g., Binongo, 2003), but knowing that a particular person uses the word "above" a lot tells us little about that person. Argamon (2006) has provided a theoretical analysis of one particular method, but in the unfamiliar and "inhuman" language of statistics, which again sheds little light on authorial language and authorial thought. By contrast, studies of gender differences in language (e.g., Coates, 2004) offer not only lists of differences, but explanations in terms of the social environment.

This is in marked contrast to some of the early (pre-computer) work in authorship analysis, which attempted to infer authorship on the basis of personality traits or psychological attributes. For example, one of the oft-suggested measures is vocabulary size, which we can easily associate with both high intelligence (a personal trait) as well as high education (a background trait). This idea can be attributed both to Simpson (1949) and Yule (1944) as well as to Talentire (1976) [which admittedly is not pre-computer]. Similarly, average word length has been often proposed [going back to De Morgan (1851)] but never successful.

Why? Why the apparent disconnect between the useful measures (such as preposition count) and meaningful measures like vocabulary richness? And in particular, why does this disconnect persist when we can find both linguistic patterns that predict personality (Argamon et al, 2005; Nowson and Oberlander, 2007) and well as medically useful linguistic diagnostics (Brown et al, 2005). We suggest two possibilities; first, that the meaningful measures proposed may not be sufficiently fine-grained, and second, that the statistical measures performed lose too much information. As an example of the

first, consider that very few words, even in high-level educated writing, exceed eight letters, meaning that "word length" is an extremely coarse-grained discretization of language. Similarly, the standard method of calculating "averages" (or even means and variances) reduces the entire data set for a given author to two numbers. Many authors have suggested (and recent findings tend to support) that multivariate analysis methods should work better for authorship attribution.

In this paper, we explore a set of multivariate analyses of well-established psycholinguistic variables. The English Lexicon Project (Balota et al, 2007) provides standardized behavioral data for a set of approximately 40,000 words, including average time for lexical decision tasks (seeing a string of characters on the screen and determining whether or not they form a word), and naming time (seeing a set of letters on the screen and naming the word they form). These are widely regarded as measures of the cognitive load involved in processing that particular word, i.e. a measure of the mental "difficulty" of that word. Following similar logic to De Morgan and Yule, we assume that some people (smarter people?) will be more comfortable using "difficult" words, and that difficulty is more appropriately measured via behavioral data than via either frequency or length.

However, rather than focusing purely on average difficulty, we apply more complex multivariate statistics to the data distribution, for example, by calculating the Kolmogorov-Smirnoff distances between the distributions, a distance that can be substantial even in instances where the means and variances of the data sets are identical. The JGAAP software package (Juola, 2009) provides many different combinations of analysis methods and preprocessing, allowing us to provide a fairly comprehensive discussion of the accuracy and usefulness of these measurements in comparison with control techniques such as simple lexical statistics.

National Digital Library of Finland: Putting the Resources of Culture, Science and Teaching at Everyone's Fingertips

Kautonen, Heli

heli.kautonen@finlit.fi

Finnish Literature Society, SKS, Finland

Sainio, Tapani

tapani.sainio@fmp.fi

The National Digital Library project, Finland

Vakkari, Mikael

mikael.vakkari@nba.fi

The National Board of Antiquities, Finland

1. Aims of the Project

The National Digital Library is one of the research, innovation, and creativity environments developed under the strategic policies of the Ministry of Education of Finland. It implements national culture and science policies by increasing the availability of the digital resources of libraries, archives and museums, by developing their long-term preservation, by establishing an important research infrastructure, and by strengthening virtual learning environments.

The National Digital Library project is also the Finnish response to the joint objectives of the European Union Member States on digitisation of cultural materials and scientific information, and their electronic availability and long-term preservation. The National Digital Library will serve as the Finnish aggregator for Europeana.

The project has four main goals:

- to build one national access point, a public interface, to digital resources of libraries, archives and museums (operational in 2011)
- to digitise and make available the most essential collections of Finnish cultural heritage organisations through the public interface
- to create sustainable solution for long-term preservation of digital cultural material (finalised plan in 2010)
- to boost national competence in the area of digitisation, online availability and accessibility and long-term preservation.

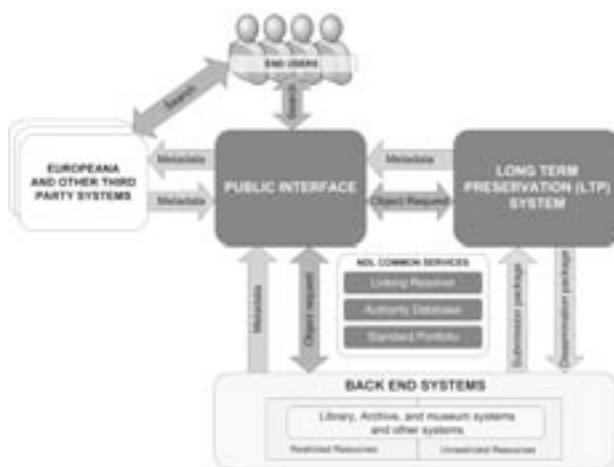


Figure 1

A total of 35 organisations participate in the project: ministries, national institutions in charge of recording and preserving cultural heritage, scientific and public libraries, archives, museums, universities, research institutes, academic associations, and representatives from other key interest groups.

The formation of consensus and a new enthusiasm for co-operation between sectors has been the key issue in an undertaking of this magnitude. On project level this has been achieved by cross-sectoral expert groups, using shared tools for project management, interaction and development, and learning from each others' practices. The central funding from the ministry has been the corner stone of this national project. The common standards defined during the project, operational public interface, and projected implementation of the long-term preservation solution should ensure the continuation of the co-operation within the Finnish memory organisation sector after the project is completed.

2. Functional principle of the Public Interface

The born-digital and digitised resources on cultural heritage, research, and teaching in Finland will be accessible to end-users through a single interface of integrated services and resources called the Public Interface. The Public Interface will be the primary access point to the resources held by libraries, museums, archives and other information suppliers in Finland. The national view (e.g. the national instance of the Public Interface) will be supplemented with, for example, customised sectoral or institutional views.

The operational principle is to continue cataloguing data and keep the digital objects in the back-end systems. Metadata will be automatically harvested to the Public Interface and normalised and indexed for easy and fast retrieval. Harvesting and services will be integrated to the Public Interface through

standardised interfaces and access to the digital objects will be provided by persistent links.

The systems architecture of the public interface separates the user interface from the back-end systems of organisations. This will facilitate customer oriented services development since more resources can be focused on user interface, services development, and integration.

The end-users' needs and expectations have played a major role during the project. Information retrieval systems currently in use are often challenging to use, and some of them require user training before information retrieval is possible. Much of the development work is based on experience gained from research on systems currently in production. The aim is to use the gained experience to develop a comprehensive and multifaceted service with high usability and fast information retrieval functionality available 24/7 anywhere. The system provides functionality for the end user to personalise the user interface to suit his/her needs.

3. Resources and Metadata

Contents of the Public Interface will be digitised or born-digital objects (images, texts, sound files, video clips, e-publications), reference data on physical objects (e.g. artefacts, books, works of art, geographical locations), or other reference data stored in databases.

The Public Interface will provide unrestricted material for all users. It will also provide restricted access materials subject to user authentication, such as licensed materials (e.g. e-journals), archive materials with restricted viewing and use, legal deposit copies, and other materials subject to copyright.

Since there are several content providers from various sectors, the metadata available will be very heterogenous. This metadata will be harvested from several different back-end systems to a centralised index of the Public Interface. The metadata will be normalised to a common internal format. This means that the Public Interface will accept any type of metadata, as long as the providing organisations have implemented standardised interfaces for harvesting.

4. Users and Usability

One of the major challenges of the project is usability. The success of the service depends greatly on its ability to meet the user expectations, which have been outlined in the requirement specifications. Usability considerations focus mostly on the user interface and its functionalities, which should expose the digital content and services within the system and enhance

their value. Furthermore, with effective user studies it will also be possible to identify potential future stakeholders and their needs.

The project has assigned a considerable amount of resources to user studies and usability design. The Usability Plan, which was realised in autumn 2009, includes five sections:

1. *Usability principles;* By following the discourse on the field of HCI as well as the accomplishments of comparable projects, the usability principles for the National Digital Library will be established.
2. *Pre-design evaluation;* Primary target groups, their roles as service users, and expectations will be described in detail. Particularly, users with vague information needs will be examined. Central use cases and projected information retrieval scenarios will be modelled. The service concept will also be tested among selected target groups.
3. *Usability evaluation;* Formal usability evaluations will be conducted: one during the piloting phase of the system, and another once the system has been deployed. All evaluations will comprise of a usability analysis and testing of the user interfaces.
4. *Tracking and evaluating actual use;* In order to provide systematic data for analysing the actual use, usage logs will be collected and recurrent user studies conducted. On-line communities of National Digital Library users will be traced, monitored, or founded.
5. *User interface design;* The project will employ a professional usability engineer for realisation of approved usability design principles, who will participate in all above mentioned usability design phases.

The results of these phases will guide the design and implementation of the Public Interface. In addition, they will have some influence on digitisation practices within organisations providing material to the portal.

5. Future Perspectives

Currently the project is preparing for the pilot phase of the Public Interface which will start after the procurement of the Public Interface software. After the pilot is complete, the Public Interface is expected to be fully implemented and go live on 2011. Plans pertaining to long-term preservation will be completed in the Summer 2010. It is, however, up to the next Finnish Government to set guidelines for its implementation and future funding.

During 2011 the NDL will transform from a project to a more sustainable organisation continuing the present work with the entire library, museum and

archive sector as an essential part of the ongoing development of the national digital infrastructure.

For more information on the National Digital Library and the outcomes, news and other deliverables of the project, see: <http://www.kdk2011.fi/en>.

Towards Digital Built Environment Studies: An Interface Design for the Study of Medieval Delhi

Keshani, Hussein

hussein.keshani@ubc.ca

University of British Columbia Okanagan, Canada

Computing technologies such as CADD, GIS or databases, are generally developed with the aims of the producers of the built environment (architects, engineers, urban planners etc.) in mind. These existing technologies tend to be adapted uncomfortably for pedagogical and research purposes. The field of built environment studies, which here refers to scholarly fields like architectural history and urban history and not practical fields like architecture, is just beginning to consider how computing technologies can be designed and employed for analytical and scholarly ends. What would software designed by practitioners of built environment studies with their aims in mind look like? Engaging with this problem is not only an opportunity to imagine a new practical tool but also to critically inquire into the aims of built environment studies and the assumptions embedded into existing built environment computing technologies. This paper presents the **Medieval Delhi Humanities Computing Research Collective's** proposal for an interface design concept that is the culmination of their attempts to analyse both their own research questions, processes and the suitability of existing technological strategies from the perspective of architectural and urban historians.

1. The Medieval Delhi Humanities Computing Research Collective

The Collective is a Canadian-led international team of historians and art and architectural historians from leading research institutions in Canada, the United Kingdom, India, and Japan with expertise in Medieval Delhi and humanities computing initiatives. Formed in 2008, the Collective is a result of the Medieval Delhi Humanities Computing Initiative funded by UBC Martha Piper Research Grant (Jan. 2008 to Sept. 2009). The Collective first met as a group in a workshop and planning session on April 2-3, 2009 in Victoria, established institutional linkages, data sharing agreements, and a common data repository, and is working together to attract additional funding. The Collective is currently

completing its work on conceptualizing researcher oriented technologies and strategies for architectural and urban historical research of Medieval Delhi.

2. Imagining Data Collages

Researchers interested in studying the built environment in a systematic way typically need to reconcile diverse forms of data – spatial, textual, and visual – and increasingly computing technologies are vital not only for storing and retrieving this information but for analyzing it as well. To be able to research built environments effectively then, a researcher-oriented digital interface and infrastructure becomes increasingly necessary. Not only does one need to build an array of databases of historical texts in multiple languages, chronologically organized photographs, maps and satellite data and other forms of information, but one needs to figure out simple productive ways to connect and interface with these various databases, integrate them with large-scale databases and design overlaying analytical tools that truly facilitate historical inquiry and collective scholarship. If planning officials, architects, tourism industries and others increasingly develop and use computing technologies with their goals in mind why should not the built environment scholarly community?

The Collective's approach treats architectural sites and urban form as a collection of visual and textual representations of varying precision across time and space each with their own interpretable contexts. For example, a site is not viewed as entirely knowable in its moment of creation but as something that evolves in form and memory and can be known only through its various representations whether they be the textual account of a 12th C court historian, the textual and pictorial accounts of a 19th C British traveller, the textual and photographic records of a 20th C Japanese archaeological team, the oral and videographic account of an Indian tourist from Mumbai, or a 21st century satellite image. These representations amount to a collage of data hence the term Data Collage. While this approach is familiar to researchers of architectural and urban history it is generally not incorporated into existing technology strategies which tend to favour virtual reconstructions or presume the stability of knowledge and a uniform level of precision for spatial and chronological information. This representational approach has important implications for how data should be structured and engaged with.

Instead of attempting to recreate a historic architectural site or region as virtual reality, Data Collages treat an architectural site or region as a collection of intersecting and conflicting representations. Ideally, a Data Collage will allow

researchers of an historic site to access all relevant three-dimensional digital models, photographs, paintings and historical textual descriptions in original and translated texts and be able to see how these various representations are interrelated chronologically and spatially and where they conflict.

References

- AlSayyad, Nezar** (1999). 'Virtual Cairo: An Urban Historian's View of Computer Simulation'. *Leonardo*. **2**: 93-100.
- Andres, Frederic, Fukami, Naoko** (2007). 'Advanced Semantic Management of Digital Resources'. *Ubicomm - International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM'07)*. Papeete, French Polynesia (Tahiti), 4-9 November 2007 V. 1, pp. 249-254.
- Ara, Matsuo** (1977). *Indo shi ni okeru Isuramu seibyō (Dargāhs in Medieval India – A Historical Study of the Shrines of Sufi Saints in Delhi with Reference to the Relationship between Religious Authority and Ruling Power)*. Tokyo: Tōkyō Daigaku Shuppankai; University of Tokyo Press.
- ARCHNET. <http://archnet.org>.
- Asami, Y., Kimura, T., Haneda , M., Fukami, N.** (2002). 'Estimation of Route and Building Sites Described in Pre-modern Travel Accounts Through Spatial Reasoning'. *Papers and Proceedings of the Geographic Information Systems Association*. **11**: 369-372.
- Thuraisingham, Bhavani M.** (2001). *Managing and mining multimedia databases*. Boca Raton, Florida: CRC Press.
- Cohen, Daniel J., et al.** (2008). 'Interchange: The Promise of Digital History'. *JAH*. **2**. <http://www.journalofamericanhistory.org/issues/952/interchange/index.html>.
- Financial Express, The** (2006). 'Rescuing our Monuments'. *The Financial Express*. **2006-02-19**. <http://www.financialexpress.com/news/rescuing-our-monuments/147737/0>.
- Fukami, Naoko** (1994). 'Studies on Muqarnas-vaulting in the Islamic Architecture: 1) the Area of central Asia: Khorasan, Khoarzum and Turan'. *Journal of the Society of Architectural Historians of Japan*. **22**: 2-36.
- Fukami, Naoko** (1996). 'Studies on Muqarnas-vaulting in the Islamic Architecture: 2) the Area of Iran: Mazandaran, Azerbaijan, Tehran, Isfahan and Yazd-Fars-kerman'. *Journal of the Society of Architectural Historians of Japan*. **25**: 23-61.
- Fukami, Naoko** (1996). 'Studies on Muqarnas-vaulting in the Islamic Architecture: 3) the Areas of Anatolia, Syria and Iraq'. *Journal of the Society of Architectural Historians of Japan*. **27**: 2-46.
- Fukami, Naoko** (1999). 'Madrasas at Isfahan: From Architectural Features and their Distributions'. *The Memoirs of the Institute for Oriental Culture*. **137**: 257-294.
- Garrido, González** (2004). 'Madinat al-zahraí: Investigación y representación, SIGraDi 2004'. *Proceedings of the 8th Iberoamerican Congress of Digital Graphics*. Porte Alegre, Brasil, 10-12 November 2004. http://cumincades.scix.net/cgi-bin/works>Show?sigradi2004_047.
- Grabar, Oleg** (1996). *The Shape of the Holy: Early Islamic Jerusalem*. Princeton: Princeton University Press.
- Hindu, The** (2007). 'IIT to prepare a 3-D database of historical sites'. *The Hindu*. **Thursday, May 31, 2007**. <http://www.hindu.com/thehindu/holnus/002200705311730.htm>.
- Husain, M.A.** (1936). *A Record of All the Qur'anic and Non-Historical Epigraphs on the Protected Monuments of the Delhi*. New Delhi: ASI, reprint 1999.
- INTACH** (2000). *Delhi: The Built Heritage—A Listing*. Delhi: Intach Delhi Chapter.
- IRCICA. <http://www.islamicarchitecturedatabase.org>.
- Keshani, Hussein** (1999). *Building Nizamuddin: A Delhi Sultanate Dargah and its Surrounding Buildings*. M.A. Thesis University of Victoria.
- Kumar, Sunil** (2008). 'Balancing Autonomy with Service: Frontier Military Commanders and their Relations with the Delhi Sultans in the 13th and 14th centuries, Presidential Address, Medieval History Section'. *Proceedings of the Punjab History Congress*. Patiala.
- Kumar, Sunil** (2008). 'The Ignored Elites: Turks, Mongols and a Persian Secretarial Class in the early Delhi Sultanate'. *Modern Asian Studies*. **1**: 45-77.
- Kumar, Sunil** (2007). *The Emergence of the Delhi Sultanate, 1192-1286*. Delhi: Permanent Black.
- Kumar, Sunil** (2007). *The Present in Delhi's Pasts*. Delhi: Three Essays Collective, Second Edition.
- Kumar, Sunil** (2000). 'Assertions of Authority: a Study of the Discursive Statements of Two Sultans of Delhi—'Ala al-Din Khalaji and Nizam al-Din Auliya'. *The Making of Indo-Persian Culture: Indian and French Studies*. Alam, Muzaffar, Delvoye, Francoise, Gaborieau, Marc (eds.). Delhi: Manohar, pp. 37-65.

- Kumar, Sunil** (2008). 'Juzjani, Minhaj-i Siraj'. *Encyclopaedia Iranica*. Leiden: E.J. Brill.
- Kumar, Sunil** (2001). 'Qutb and Modern Memory'. *Partitions of Memory*. Kaul, Suvir (ed.). Delhi: Permanent Black, pp. 140-182.
- Kumar, Sunil** (2006). 'Service, Status and Military Slavery in the Delhi Sultanate of the thirteenth and early fourteenth centuries'. *Slavery and South Asian History*. Eaton, Richard, Chatterjee, Indrani (eds.). Bloomington: Indiana University Press.
- Kumar, Sunil** (2008). 'Trans-regional Contacts and Relationships in Islam's Eastern frontier: the Delhi Sultanate in the thirteenth and fourteenth Centuries'. *Annales Islamologiques*.
- Kumar, Sunil** (1999). 'Perceiving 'your' Landscape: Neighbourhood Settlements and the Hauz-i Rānī'. *Perceiving Landscape*. Layton, R., Ucko, P. (eds.). London: Routledge and Kegan Paul, pp. 159-174.
- Kumar, Sunil** (1994). 'When Slaves were nobles: The Shamsi bandagan in the early Delhi Sultanate'. *Studies in History*. 23-52.
- Matsubara, Kosuke, Fukami, Naoko, Arai, Yuji, Imamura, Fumiaki, Iitsuka Mayumi, Yamada Eri** (2005). 'A study on preservation for the old city of Fez, Morocco-From the Viewpoint of the Separation of Public and Private'. *Journal of Housing Research Foundation*. 31: 113-124.
- Invitation of Proposal for Procurement of High-Resolution Imagery for Delhi*. <http://www.mcdonl ine.gov.in/>.
- Mohan, Madan** (2004). 'Spatial Data Modeling in GIS for Historical Restoration and Conservation of Cultural Heritage of Seven Cities Of Delhi'. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*. vol. 35, part 5: 896-902.
- Mohan, Madan** (2004). 'Historical Information System for Surveying Monuments and Spatial Data Modeling for Conservation of Cultural Heritage in Delhi'. *Technical Papers, FIG Working Week, Intercontinental Athenaeum Athens*. Athens, Greece, 22-27 May 2004.
- Okabe, Atsuyuki** (2006). *GIS-Based Studies in the Humanities and Social Sciences*. Boca Raton: Taylor Francis.
- Okabe, Atsuyuki** (2004). *Islamic Area Studies With Geographical Information Systems*. New York: Routledge.
- Page, J.A., Hasan. Zafar** (1916). *Monuments of Delhi*. Delhi: ASI, reprint 2004.
- Soja, Edward** (1989). *Postmodern Geographies*. London: Verso.
- Soja, Edward** (2003). 'Writing the City Spatially'. *City: analysis of urban trends, culture, theory, policy, action*. 3: 269-281.
- Sokhi, B. S.** (1992). 'Spotting historical monuments and sites: Mapping of historical monuments and sites of Delhi using SPOT satellite image'. *Journal of the Indian Society of Remote Sensing*. 2-3: 65-71.
- Properties inscribed on the World Heritage List*. <http://whc.unesco.org/en/statesparties/in>.
- Verma, Richi** (2008). 'City's unprotected monuments dying a slow death'. *The Times of India*. 9 Jun 2008. http://timesofindia.indiatimes.com/C ities/Delhi/Citys_unprotected_monuments_dy ing_a_slow_death/rssarticleshow/3112362.cm s.
- Welch, Anthony** (1996). 'Gardens that Babur Did Not Like: Landscape, Water, and Architecture for the Sultans of Delhi'. *Mughal Gardens: Sources, Places, Representations, and Prospects*. Wescoat, J. L., Wolschke-Bulmahn, J. (eds.). Washington DC: Dumbarton Oaks Research Library, Harvard University, pp. 59-93.
- Welch, Anthony** (1984). 'Qur'an and Tomb, the religious epigraphs of two early Sultanate tombs in Delhi'. *Indian Epigraphy, its Bearing on Art History*. Asher, F., Gai, S. (eds.). Washington, D.C.: Smithsonian Institution, pp. 247-257.
- Welch, Anthony, Keshani, Hussein, Bain, Alexandra** (2002). 'Epigraphs, Scripture, and Architecture'. *Muqarnas, an Annual on the Visual Culture of the Islamic World, The Aga Khan Program at Harvard and M.I.T.* XIX: 12-43.
- Welch, Anthony** (1997). 'The Shrine of the Holy Footprint in Delhi'. *Muqarnas, an Annual on the Visual Culture of the Islamic World, The Aga Khan Program at Harvard and M.I.T.* XIV: 166-178.
- Welch, Anthony** (1996). 'A Medieval Muslim Center of Learning in India: the Hauz Khas Madrasa in Delhi'. *Muqarnas, an Annual on the Visual Culture of the Islamic World, The Aga Khan Program at Harvard and M.I.T.* XIII: 165-190.
- Welch, Anthony** (1993). 'Architectural Patronage and the Past: The Tughluq Sultans of Delhi'. *Muqarnas, an Annual on the Visual Culture of the Islamic World, The Aga Khan Program at Harvard and M.I.T.* X: 311-322.
- Welch, Anthony** (1985). 'Hydraulic Architecture in Medieval India: The Tughluqs'. *Environmental Design, Journal of the Islamic Environmental Design Research Centre*. 2: 74-81.
- Welch, Anthony, Crane, Howard** (1984). 'The Tughluqs: Master Builders of the Delhi Sultanate'.

Muqarnas, an Annual on the Visual Culture of the Islamic World, The Aga Khan Program at Harvard and M.I.T. I: 123-166.

Welch, Anthony (2005). 'The Qutb Minar'. *Encyclopedia of Medieval Islamic Civilization*. London: Routledge.

Welch, Anthony (1997). 'Tughlukids: Architecture'. *Encyclopaedia of Islam*. Leiden: E.J. Brill.

Yamamoto, T., Ara, M. and Tsukinowa, T. (1968-70). *Delhi: Architectural Remains of the Sultanate Period (in Japanese)*. Indo Shiseki Chōsa Dan hen. 3 vols.

Propp Revisited: Integration of Linguistic Markup into Structured Content Descriptors of Tales

Lendvai, Piroska

piroska@nytud.hu

Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary

Declerck, Thierry

declerck@dfki.de

Language Technology Lab, DFKI GmbH,
Saarbrücken, Germany

Darányi, Sándor

Sandor.Daranyi@hb.se

Swedish School of Library and Information Science,
University College Boras/Göteborg University,
Sweden

Malec, Scott

malec@andrew.cmu.edu

Carnegie Mellon University, USA

Metadata that serve as semantic markup, such as conceptual categories that describe the macrostructure of a plot in terms of actors and their mutual relationships, actions, and their ingredients annotated in folk narratives, are important additional resources of digital humanities research. Traditionally originating in structural analysis, in fairy tales they are called functions (Propp, 1968), whereas in myths – mythemes (Lévi-Strauss, 1955); a related, overarching type of content metadata is a folklore motif (Uther, 2004; Jason, 2000).

In his influential study, Propp treated a corpus of tales in Afanas'ev's collection (Afanas'ev, 1945), establishing basic recurrent units of the plot ('functions'), such as *Villainy*, *Liquidation of misfortune*, *Reward*, or *Test of Hero*, and the combinations and sequences of elements employed to arrange them into moves.¹ His aim was to describe the DNA-like structure of the magic tale sub-genre as a novel way to provide comparisons. As a start along the way to developing a story grammar, the Proppian model is relatively straightforward to formalize for computational semantic annotation, analysis, and generation of fairy tales. Our study describes an effort towards creating a comprehensive XML markup of fairy tales following Propp's functions, by an approach that integrates functional text annotation

with grammatical markup in order to be used across text types, genres and languages.

The Proppian fairy tale Markup Language (PftML) (Malec, 2001) is an annotation scheme that enables narrative function segmentation, based on hierarchically ordered textual content objects. We propose to extend PftML so that the scheme would additionally rely on linguistic information for the segmentation of texts into Proppian functions. Textual variation is an important phenomenon in folklore, it is thus beneficial to explicitly represent linguistic elements in computational resources that draw on this genre; current international initiatives also actively promote and aim to technically facilitate such integrated and standardized linguistic resources. We describe why and how explicit representation of grammatical phenomena in literary models can provide interdisciplinary benefits for the digital humanities research community.

In two related fields of activities, we address the above as part of our ongoing activities in the CLARIN² and AMICUS³ projects. CLARIN aims to contribute to humanities research by creating and recommending effective workflows using natural language processing tools and digital resources in scenarios where text-based research is conducted by humanities or social sciences scholars. AMICUS is interested in motif identification, in order to gain insight into higher-order correlations of functions and other content units in texts from the cultural heritage and scientific discourse domains. We expect significant synergies from their interaction with the PftML prototype.

1. Proppian fairy tale Markup Language (PftML)

Creating PftML was based on the insight that Propp's functions – organized in tables to categorize his observations – were analogous to metadata, and as such renderable by hierarchically arranged elements in eXtensible Markup Language (XML) documents. A tale consists of one or more moves and on a lower level of functions which are modeled as elements. Function elements themselves have XML attributes that allow for the efficient extraction of data from the text using XQuery from within a native XML database. The embedded structure of Proppian functions as represented by PftML markup is illustrated in Fig. 1 by an annotated excerpt from the English translation of the Russian fairy tale *The Swan-Geese*.

Note that Proppian functions are applied to relatively long, semantically coarse-grained textual chunks, i.e. sentences, but linguistic elements that convey a function actually encompass a shorter sequence of

words; e.g. contrary to the markup in the example, both *Command* and *Execution* only pertain to linguistic units smaller than full sentences.

```

<Corpus>
<Folktales Title="The Swan-Geese" AT="480"
  NewAfanasievEditionNumber="113"
  ProppConformity="Yes">
<Move>
<Preparation>
  <InitialSituation> Once upon a time a man
  and a woman lived with their daughter and small
  son. </InitialSituation>
  <CommandExecution>
    <Command subtype="Interdiction">
      "Dearest daughter," said the mother, "we are
      going to work. Look after your brother! Don't
      go out of the yard, be a good girl, and we'll
      buy you a handkerchief." </Command>
    <Execution subtype="Violated">
      The
      father and mother went off to work, and the
      daughter soon enough forgot what they had told
      her. She put her little brother on the grass
      under a window and ran into the yard, where she
      played and got completely carried away having
      fun.</Execution>
  </CommandExecution>
</Preparation>
<Villainy subtype="Kidnapping">
  In swooped the
  swan-geese, snatched up the little boy, and
  flew away with him. </Villainy>
<ConsentToCounteraction>
  When the girl came
  back inside, her brother was missing! "Oh
  no!" she cried. She dashed here and there, but
  there was no sign of him. She called for him,
  cried, and wailed how angry mother and father
  would be, but her brother did not answer. </
  ConsentToCounteraction>

```

2. Integration of PftML with linguistic annotation

We propose to combine PftML with a stand-off, multi-layered linguistic markup scheme to ensure modularity and reusability of linguistic information associated with textual elements, supporting interoperability of fairy tales annotation in different languages and versions. As seen in Fig. 1, PftML is interleaving the Proppian annotation with the text. This in-line annotation strategy has some drawbacks: a text can hardly be annotated in fine-grained manner without losing readability, or with information originating from different sources e.g. indicating different views on narrative functions.

Stand-off annotation strategy, following the standardization initiatives for language resources conducted within ISO,⁴ stores annotation separately from the original text, linking these by referencing mechanisms. We adopt the ISO multi-layered annotation strategy, representing linguistic information on the following levels: segmentation of the text in tokens; morpho-syntactic properties of the tokens; phrasal constituents; grammatical

dependencies; semantic relations (e.g. temporal, co-referential), cf. (Ide and Romary, 2006).

We illustrate how the linguistic annotation layers can be combined with the PftML annotation in one stand-off annotation file, showing here only the morphosyntactic and constituency annotation, as they are applied to the first five tokens of the sub-sentence annotated with the 'Violated Execution' function in Fig. 1. In the morphosyntactic annotation, the value of the `tokenID` of the 12th word is pointing to the original data (e.g., *daughter* is the 12th token in the text).⁵

```
<wordForms>
    <W ID="w11" POS="ART" LEMMA="the" MORPH="Sg"
    tokenID="t11">the</W>
    <W ID="w12" POS="NN" LEMMA="daughter"
    MORPH="Sg" tokenID="t12">daughter</W>
    <W ID="w13" POS="ADV" LEMMA="soon"
    tokenID="t13">soon</W>
    <W ID="w14" POS="ADV" LEMMA="enough"
    tokenID="t14">enough</W>
    <W ID="w15" POS="VFIN" LEMMA="forget"
    MORPH="Past" tokenID="t15">forgot</W>
    ...
</wordForms>
```

In the constituency annotation level displayed below, words are grouped into syntactic constituents (e.g. the nominal phrase *the daughter*). The span of constituents is marked by the value of the features `from` and `to`, which are pointing to the previous morpho-syntactic annotation layer.

```
<phrases>
    <phrase id="p4" from="w11" to="w12"
    type="NP">the daughter</phrase>
    <phrase id="p5" from="w13" to="w14"
    type="ADVP">soon enough</phrase>
    <phrase id="p6" from="w15" to="w15"
    type="VG">forgot</phrase>
    <phrase id="p7" from="w16" to="w20"
    type="REL_COMP">what they had told her
    </phrase> ...
</phrases>
```

PftML and (for example) word-level annotation can be combined in one stand-off XML element, where each specific PftML annotation receives a span of textual segments associated with it:

```
<Execution subtype="Violated" id="e1"
    inv_id="Command1" from="w11" to="w21"> </
Execution>
```

The values `w11` and `w21` are used for defining a region of the text for which the Propp function holds; `Command1` refers to the *Interdiction* function label used earlier in the text.⁶

3. Benefits for humanities research

The integrated annotation scheme enables narrative segmentation enhanced by additional information about the linguistic entities that constitute a given function. A folklore researcher might be interested in which natural language expressions correspond to which narrative function: in the `<Execution subtype="Violated">` example, *forgot* can be an indicator of this function. In fact, it is also relevant to signal that *forgot* is a verb, and to reduce the strings *forgets*, *forgot*, *forgotten* to one lemma (i.e. base form), so that all morphological forms are retrieved when any of these variants is queried.

Navigating through the different types of IDs included in the multilayered annotation, a researcher can obtain statistics over linguistic properties of fairy tales. For example, the grammatical subject of a function can be extracted, e.g. to see which characters participate in commands and their violation. Note that if – according to the current scheme – the narrative function boundaries are imprecise, the `<Execution subtype="Violated">` function in our example sentence would incorrectly contain two grammatical – and three semantic – subjects (*father* and *mother*, and *daughter*).

Linguistic information will enable detecting functions that refer to each other, as syntax and semantics of sentence pairs in such relations mirror – at least partly – each other, e.g. *Don't go out of the yard* and *ran onto the street*. Detecting cross-reference in turn contributes to identifying a function's core elements, which is a crucial step in understanding the linguistic vehicles by which motifs operate and the degree of variation and optionality they allow.

4. Concluding remarks

Since the content descriptors in PftML might pertain to textual material on the supra- or subsentential level, there is a need to investigate the mechanisms underlying the assignment of a function to a span of words. We propose to tackle this issue based on linguistic analysis, hypothesizing that boundaries of certain linguistic objects overlap with boundaries of Proppian functions. A direct consequence of more precise segmentation of functions is that linguistic characterization, retrieval, and further computational processing of texts from the folktale genre will improve, and facilitate detecting higher-level, domain-specific cognitive phenomena. It would also become feasible to detect from corpus evidence if there exist additional functions beyond Propp's scheme.

Integration along the above lines with ontological resources of fairy tales is described in a separate study by us (Lendvai et al., 2010). We expect from our strategy – applied to tales in different versions in different languages – to lead to the generation of a multilingual ontology of folktale content descriptors, which would be extending the efforts of the MONNET project,⁷ originally focussing on financial and governmental issues. In future work we plan to address embedding our annotation work into the TEI framework,⁸ and extend the ISO strategy on using well-defined data categories for linguistic annotation labels⁹ to those of functions corresponding to PftML labels, to facilitate porting our approach to other literary genres.

6. We started to implement this work within the D-SPIN project (see <http://www.sfs.uni-tuebingen.de/dspin>), which is the German complementary project to CLARIN.
7. Multilingual ONtologies for NETworked Knowledge, see http://cordis.europa.eu/fp7/ict/languagetechnologies/project-monnet_en.html
8. <http://www.tei-c.org/index.xml>
9. see <http://www.isocat.org/>

References

- Afanas'ev, A.** (1945). *Russian fairy tales*. New York: Pantheon Books.
- Ide, N. and Romary, L.** 'Representing linguistic corpora and their annotations'. *Proc. of LREC*. 2006.
- Jason, H.** (2000). *Motif, type and genre. A manual for compilation of indices and a bibliography of indices and indexing*. Helsinki: Academia Scientiarum Fennica.
- Lendvai, P., Declerck, T., Darányi, S., Gervás, P., Hervás, R., Malec, S., and Peinado, F..** 'Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case'. *In Proc. of LREC*. 2010.
- Lévi-Strauss, C.** (1955). 'The structural study of myth'. *Journal of American Folklore*. **68**: 428–444.
- Malec, S. A.** 'Proppian structural analysis and XML modeling'. *In Proc. of CLiP*. 2001.
- Propp, V. J.** (1968). *Morphology of the folktale*. Austin: University of Texas Press.
- Uther, H. J.** (2004). *The types of international folktales: a classification and bibliography. Based on the system of Antti Aarne and Stith Thompson*. Helsinki: Academia Scientiarum Fennica.

Notes

1. The full list of functions is available at <http://clover.slavic.pitt.edu/sam/propp/praxis/features.html>
2. <http://www.clarin.eu>
3. <http://ilk.uvt.nl/amicus>
4. <http://www.tc37sc4.org>
5. Note that the original string is normally not present in these layers but is displayed in annotation examples for readability's sake.

Extracting domain knowledge from tables of contents

Harald Lüngen

luengen@uni-giessen.de

Justus-Liebig-Universität Gießen

Henning Lobin

henning.lobin@germanistik.uni-giessen.de

Justus-Liebig-Universität Gießen

1. Introduction

Knowledge in textual form is always presented as visually and hierarchically structured units of text, which is particularly true in the case of academic texts. One research hypothesis of the ongoing project *Knowledge ordering in texts—text structure and structure visualisations as sources of natural ontologies*¹ is that the textual structure of academic texts effectively mirrors essential parts of the knowledge structure that is built up in the text. The structuring of a modern dissertation thesis (e.g. in the form of an automatically generated table of contents - tocs), for example, represents a compromise between requirements of the text type and the methodological and conceptual structure of its subject-matter. The aim of the project is to examine how visual-hierarchical structuring systems are constructed, how knowledge structures are encoded in them, and how they can be exploited to automatically derive ontological knowledge for navigation, archiving, or search tasks. The idea to extract domain concepts and semantic relations mainly from the structural and linguistic information gathered from tables of contents represents a novel approach to ontology learning.

2. Data and annotations

In the present phase, we examine German academic text books, in later phases, dissertations, research articles and historical scientific texts will also be taken into account. A corpus of digital versions of 32 text books from 12 different academic disciplines has been compiled,² the textual content and an XML document structure markup was subsequently extracted (e.g. using the Adobe Pro software). Using a series of XSLT style sheets, the initial XML was converted to XML encoding of the document structure according to the TEI P5 guidelines. At the same time, the texts were annotated with morphological analyses and phrase

chunking markup using the Tree Tagger and Chunker software from Stuttgart University³ and converted to a suitable XML representation, and also with dependency-syntactic analysis using the Machinese Syntax Parser by Connexor Oy,⁴ resulting in XML markup as well. Further linguistic annotation levels (such as domain terms and lexical-semantic relations) will be added and combined in XStandoff documents representing multi-layer annotations that can be queried using XML standards and tools as described in Stührenberg & Jettka (2009).

Presently, all available annotation layers are stored in an eXist native XML database⁵ and are queried using the Oxygen XML editor⁶ as a database client.

The corpus infrastructure is used to explore the document applying the method of toc fragment analysis as described in the following section, and to implement functions for concept extraction and semantic relation analysis.

2.1. Analysis of toc fragments

Our method of analysing toc fragments consists of the following steps:

1. Identification of a toc fragment
2. Representation of the fragment meaning as a MultiNet
3. Identification of the configuration of elements on different structural levels that induce the fragment meaning
4. Hypothesis about the generalisation of a toc fragment, a structuring schema
5. Corpus research to verify the generalisation hypothesis

```
[booktitle] Einführung Pädagogik
[5.] Ausgewählte Subdisziplinen und
Fachrichtungen
[5.1.] Literatur
[5.2.] Erlebnispädagogik
[5.2.1.] Begrifflichkeit
[5.2.1.1.] Erlebnis als prioritäre Kategorie
[5.2.2.] Historie
[5.2.3.] Theoretische Fundierungen und
Menschenbilder
[5.2.4.] Ziele und Funktionen der
Erlebnispädagogik
[5.2.4.1.] Ziele
[5.2.4.2.] Subjektbezogene Funktionen und
mögliche Wirkungsweisen
[5.2.4.3.] Gesellschaftliche Funktionen
[5.2.5.] Merkmale und Modelle der
Erlebnispädagogik
[5.2.6.] Beispiele erlebnispädagogischer
Angebote
[5.2.6.1.] Outward Bound-Konzeption
[5.2.6.2.] Outdoor Management Development
[5.2.7.] Kritikpunkte
```

[5.2.8.] Einführungsliteratur (zum Weiterlesen)
[5.3.] Erwachsenenbildung
[5.3.1.] Begriffsklärung
[5.3.2.] Geschichtliche Entwicklung
[5.3.3.] Struktur und Funktionsperspektiven in der Erwachsenenbildung
[5.3.4.] Theoretische Orientierungen der Erwachsenenbildung
[5.3.5.] Forschungsfelder
[5.3.6.] Einführungsliteratur (zum Weiterlesen)
[5.4.] Gesundheitspädagogik

Figure 1: Section from the table of contents of Raithel et al. (2007)

Consider the section of the generated table of contents of the text book *Einführung Pädagogik* by Raithel et al. (2007) shown in Figure 1. By choosing the heading 5. *Ausgewählte Subdisziplinen und Fachbereiche* and its immediately superordinated heading (in this case the title of the book) as well as its immediately subordinated headings, we arrive at the toc fragment (or “window”) shown in Figure 2. In the toc fragment, four terms from the domain are contained, *Pädagogik*, *Erlebnispädagogik*, *Erwachsenenbildung*, and *Gesundheitspädagogik*.⁷ The terms identification component must distinguish such expressions denoting domain-specific concepts from relational nouns commonly found in academic and scientific discourse (such as *Einführung*, *Subdisziplin* and *Fachrichtung*) and from terms denoting text-type structural categories of academic texts such as *Literatur*.

We employ the semantic network approach *Multilayered Extended Semantic Networks* (acronym: MultiNets) by Helbig (2006) to represent the domain concepts and semantic relations between them

[booktitle] Einführung in die Pädagogik
[5.] Ausgewählte Subdisziplinen und Fachrichtungen
[5.1.] Literatur
[5.2.] Erlebnispädagogik
[5.3.] Erwachsenenbildung
[5.4.] Gesundheitspädagogik

Figure 2: Toc fragment

expressed in a toc fragment. The MultiNet approach is a fully-fledged semantic theory and provides a rich and consistent inventory of semantic entity types, features, relations and functions, and has been previously employed in the syntactic-semantic analysis components of QA systems (Hartrumpf 2005). Using the graphical MWR editor for designing MultiNets,⁸ we represent the semantics of the above toc fragment as shown in Figure 3.

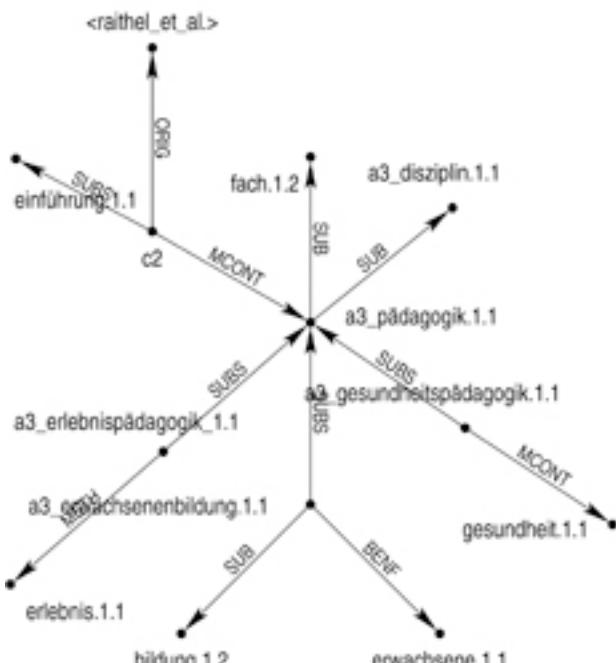


Figure 3: MultiNet analysis of toc fragment

In the semantic network in Figure 3, the concepts *a3_erlebnispädagogik.1.1*, *a3_ergebnissenbildung.1.1*, and *a3_gesundheitspädagogik.1.1* are related to *a3_pädagogik.1.1* by the SUBS relation denoting subordination of (abstract) situations; *a3_pädagogik.1.1* is in turn related to *fach.1.2* and *a3_disziplin.1.1* by the SUB relation denoting the subordination of concepts representing objects.⁹ Furthermore, the semantic decompositions of the three compounds are analysed using the relations MCONT (mental or informational content), BENF (beneficiary) and METH (method), and the relation between the concept *c2* representing the textbook as such and *a3_pädagogik.1.1* is specified as MCONT, the relation between *c2* and its authors as ORIG (mental of informational origin, cf. Helbig 2006).

On account of this analysis, the following hypothesis is formed:

Given a potential structuring schema, consisting of an initial expression N, and an expression N-1 related to N by a *heading_of* relation on the document structure level, and an expression N+1 to which N is related by the *heading_of* relation on the document structure level (cf. Figure 4), if



Figure 4: Toc Schema

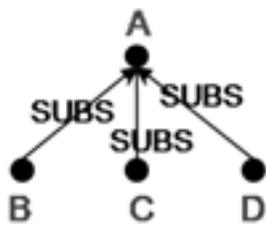


Figure 5: MultiNet Schema

- N contains the lexeme Subdisziplin or a synonym on the lexical level
- and N-1 contains the domain concept A
- and N+1 contains the domain concept B,

then, by multiple application, construct a MultiNet-Schema as represented by the graph in Figure 5.

```

<result doc="schruender-
lenzen_schriftspracherwerb_2007" docID="i72">
    <head level="n-1">9. Schwierigkeiten des
    Schriftspracherwerbs rechtzeitig erkennen und
    gezielt helfen</head>
    <head level="n">9.2 Zentrale
    Wahrnehmungsbereiche und ihr Risikopotential</
    head>
    <head level="n+1">9.2.1 Visuelle
    Wahrnehmung</head>
    <head level="n+1">9.2.2 Auditive
    Wahrnehmung</head>
</result>
<result
doc="brosius_kommunikationsforschung_2008"
docID="i137">
    <head level="n-1">8. Kapitel:
    Inhaltsanalyse I: Grundlagen</head>
    <head level="n">8.4 Anwendungsgebiete und
    typische Fragestellungen</head>
    <head level="n+1">8.4.1 Inhaltsanalysen auf
    dem Feld der politischen Kommunikation</head>
    <head level="n+1">8.4.2 Inhaltsanalysen in
    der Gewaltforschung</head>
    <head level="n+1">8.4.3 Inhaltsanalysen in
    der Minderheitenforschung</head>
</result>
<result doc="raithel_paedagogik_2007"
docID="i180">
    <head level="n-1">BOOKTITLE: Einführung
    Pädagogik</head>
    <head level="n">D Ausgewählte
    Subdisziplinen und Fachrichtungen</head>
    <head level="n+1"> Literatur</head>
    <head level="n+1">Erlebnispädagogik</head>
    <head level="n+1">Erwachsenenbildung</
    head>
    ...
  
```

Figure 6: Query Result Document

The Hypothesis is verified by formulating the potential structuring schema as a query to the corpus using the XQuery query language. The query result document then contains a set of toc fragments that can now be inspected as to whether their semantics conform to the hypothesis or not, leading to a small statistic about the validity of the hypothesis.

Sometimes the inspection may also lead to a modification of the original query. In the first result fragment in Figure 6, for instance, the superordinate concept *Wahrnehmung* is not contained in N-1, but as the compound modifier of *Bereich* (a synonym of *Subdisziplin*).¹⁰

In this example it becomes clear that analyses on the morphological and lexical-semantic level interact with the analyses of the structuring information in that both levels provide conditions or constraints when building the semantic analysis of a toc fragment. Our corpus infrastructure is designed such that information from multiple linguistic and structural levels can be taken into account.

2.2. Conclusion

We presently inventorise sets of complex conditions connecting a structuring schema with a MultiNet Schema as *constructions* in the sense of Construction Grammar (CxG). Construction Grammar (Kay 1995, Östman & Fried 2004) is a theory of grammar which is not based on phrase structure rules operating on lexical elements, but as combinations of constructions in which form schemata are associated with meaning schemata and is therefore appropriate for the description task at hand. The inventory of constructions will then be employed in ontology learning, particularly for the task of automatically extracting domain concepts and semantic relations between them. Constructions describing document structuring schemata as described above play a role similar to the lexico-syntactic "Hearst Patterns" described in Hearst (1992), which have been employed for extracting semantic relations from running text.

References

- Brosius, Hans-Bernd, Koschel, Friederike, Haas, Alexander** (2008). *Soziologie. Methoden der empirischen Kommunikationsforschung.* 4. Aufl. Wiesbaden.
- Glöckner, Ingo, Hartrumpf, Sven, Helbig, Hermann, Leveling, Johannes, Osswald, Rainer** (2007). 'Automatic semantic analysis for NLP applications'. *Zeitschrift für Sprachwissenschaft.* Jg. 26, H. 2: 241–266.
- Hartrumpf, Sven** (2005). 'University of Hagen at QA@CLEF 2005: Extending knowledge and deepening linguistic processing for question answering.' *Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop.* Peters, Carol (ed.). Vienna: Centromedia.

- Hearst, Matti A.** (1992). 'Automatic acquisition of hyponyms from large text corpora'. *Proceedings of the 14th International Conference on Computational Linguistics*.
- Helbig, Hermann** (2006). *Knowledge Representation and the Semantics of Natural Language*. Cognitive Technologies. Heidelberg: Springer
- Kay, Paul** (1995). 'Construction Grammar'. *Handbook of Pragmatics Manual*. Verschueren, Jef, Östman, Jan-Ola, Blommaert, Jan (eds.). Amsterdam: John Benjamins, pp. 171–177.
- Östman, Jan-Ola, Fried, Mirjam (eds.)** (2004). *Construction Grammars: Cognitive grounding and theoretical extensions*. Amsterdam: John Benjamins.
- Raithel, Jürgen, Dollinger, Bernd, Hörmann, Georg** (2007). *Einführung in die Pädagogik. Begriff - Strömungen - Klassiker - Fachrichtungen*. 2. Aufl.. VS Verlag für Sozialwissenschaften. Wiesbaden: Springer
- Schmid, Helmut** (1994). 'Probabilistic Part-of-Speech Tagging using Decision Trees'. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.
- Schründler-Lenzen, Agi** (2007). *Schriftspracherwerb. Bausteine professio-nellen Handlungswissens*. 2. Aufl. VS Verlag für Sozialwissenschaften. Wiesbaden: Springer
- Stührenberg, Maik, Jettka, Daniel** (2009). 'A toolkit for multi-dimensional markup: the development of SGF to XStandoff'. *Proceedings of Balisage: The Markup Conference 2009*. Balisage Series on Markup Technologies. 3 vols.
- Tapanainen, Pasi, Järvinen, Timo** (1997). 'A non-projective dependency parser'. *Proceedings of the fifth conference on Applied natural language processing*. Washington D.C.
-
- Notes**
1. funded within the framework of LOEWE, the excellence initiative of the state of Hesse, as part of the *LOEWE-Schwerpunkt Kulturtechniken und ihre Medialisierung*, cf. <http://www.zmi.uni-giessen.de/projekte/projekt-36.html>.
 2. We would like to thank the publishers *Facultas, Haupt, Narr/Francke/Attempto, Springer, UTB, Vandenhoeck & Ruprecht*, and *Wissenschaftliche Buchgesellschaft* for kindly making available digital versions of textbooks for us.
 3. <http://www.ims.uni-stuttgart.de/projekte/corpora/TreeTagger/>
 4. <http://www.connexor.eu/>
 5. <http://exist-db.org/>
 6. <http://www.oxygenxml.com/>
 7. We consider terms to be linguistic expressions that refer to domain concepts.
 8. which was kindly made available for us by Professor Helbig's group in Hagen.
 9. An a3_ prefix in a concept name indicates that the concept was not found in the required reading in the semantic lexicon HagenLex (Glöckner et al. 2007) which is consulted by the MWR tool.
 10. Other titles from our corpus cited in Figure 6 are Brosius (2008), and Schründler-Lenzen (2008).

Museums of the virtual future

Marie-Madeleine Martinet

marie-madeleine.martinet@paris-sorbonne.fr
Department of English, Université Paris IV-Sorbonne, France

Liliane Gallet-Blanchard

liliane.gallet-blanchard@paris-sorbonne.fr
Department of International Business and Languages, Université Paris IV-Sorbonne, France

The poster will argue that a retrospective museum-like exhibition of digital media leads to further developments in the field of ‘visualisation’: the study of the past offers new opportunities to emerging technologies. The poster is based on an exhibition entitled ‘Is the virtual real?’ which took place at the University of Paris-Sorbonne in October 2009, organized by the Research Centre CATI (Cultures Anglophones et Technologies de l’Information <http://www.cati.paris-sorbonne.fr>), and on the work in progress of the digital preservation of the exhibition.

It will be supported by computer demonstrations showing 1) the time-perspective issues in a retrospect on the history of Virtual Reality: examples of the variety of the original documents 2) the interaction between different areas of expertise in IT related to visualisation, and between history and IT practice, during the exhibition 3) the preservation of the exhibition on digital records as a source of new projects.

1. A museum display on the history of virtual reality: from simulation to image

1.1. From the past to the present: bringing precursors together

A retrospect in a museum of science will have to show that present-day IT applications are a convergence of many technologies of the past, eventually leading to simulation.

The first micro of the 1970s, the Micral N (1973) was made for a laboratory in agronomy studying evaporation phenomena. Still earlier, an analogue machine dating back to the 1960s, the EAI 580 (1963), before the digital age, was used to calculate possible options and variations for industrial

processes such as crane counterweights and plane landing without visibility; it was later used in research laboratories. The idea of ‘simulation’ existed long before it converged with digital technology, which in early days lacked the necessary computing power, and took over only when its computing speed made it appropriate for simulation. Image or simulation technologies were first distinct from computing, and a retrospect has to show the separate strands.

1.2. From the future to the past: VR and interactivity

The present computing power gives a new approach to history, by allowing us to ‘experiment’ with the past. A hypothesis on the construction of the Egyptian pyramid can be tested with a CAD programme by an architect – a version of a present-day approach to industrial history where the expertise of professionals is required to assess the plausibility of hypotheses concerning industrial processes of the past.

VR programmes meant to reconstruct the architecture of past centuries – Italian Renaissance buildings or the Georgian city of Bath – use the same CAD tools as those employed nowadays in architects’ practices to plan for future buildings. A case in point is the Renaissance theatre which was published as a woodcut in a 16th century book, then, thanks to present-day architectural software, became a 3D virtual model, which in turn served as a plan to build a real wooden model; passing from 2D to 3D enables the viewer to visualize the gradual distortion of the stage set as the spectators sit farther from the seat of honour.

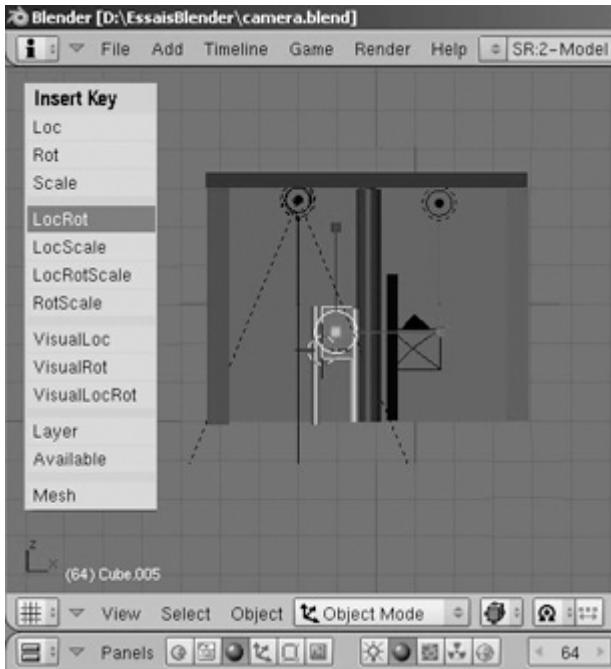
The experimental approach also concerns the public, who can navigate interactively. Examples of games exemplify the exchange between entertainment and professional simulation, such as flight simulators which may be used for both; and game engines may be used for architectural or historical simulations.



Three views of the exhibition: the nave (above) with the display of the early simulation computers lent by the Fédération des Équipes Bull, leading to screens showing films on applications of VR in medicine and in the history of architecture (top right), and to interactive programmes on famous landscapes (right): Prior Park, by CASA, Bath.

1.3. Visualisation between image and IT

3D technology, interpretation and practice



Technical practice gives a better understanding of an art. The visitors in the exhibition could experiment the creation of 3D graphics themselves on a computer at their disposal: practice on present-day software enables the visitors to visualize the interaction between geometry, technology and aesthetics that have made the history of VR – and of architecture. They experience a new view of art history, starting from cylinders and half-spheres to produce an image of Classical rotundas with circular colonnades and domes.

Various modes of presentation contextualize this combination of VR programmes running on machines, both professional finished projects and visitors' experiments: they need to be supported by explanatory panels on the history and practice of CAD and VR, placing it in a historical context – 3D modelling dating back to Renaissance perspective. The panels also have to explain the main notions of 3D graphics: the creation of primitive forms associating geometry and art, the use of textures, lighting and perspective through the 'camera eye', and eventually animation. They show how the basic techniques of geometrical coordinates experimented by the visitors underlie, after much elaboration, the artistic finished products of VR that they can admire in the exhibition.

Virtual worlds such as Second Life, where the computing power and network capabilities required can be used for artistic purposes such as a live

jazz performance broadcast internationally, also exemplify the interactive possibilities of VR.

Digital image aesthetics and the historical perspective

Films on the history of VR are shown, and here again media technology has to be harnessed so as to suggest the historical perspective, based on the practices of digital visualisation aesthetics; this involves cooperation between historians of IT and digital media technicians. Documentaries dating back to the early 1990 differ in pixel definition from cutting-edge trailers of new programmes, yet they have to appear on the same screen; videos on VR in medicine and videos on architecture are focused differently. The resulting films have to show both the similarity in techniques and the differences in purpose between documentaries in these various areas. A solution is to present the early films in a smaller format which makes pixelisation less visible, and surround them with a frame to recover the format of the larger films; the frame effect will in addition give them an old-fashioned air which will suggest the time perspective.

The museology of IT

The layout of the exhibition is meant to relate the various IT programmes presented. An option was to take advantage of the exhibition space – in our case, a linear two-nave 15th century building, in which a circuit was organized so as to trace the history of VR from its beginning to present-day supercomputers. The lighting is also of great importance, implying cooperation between the electricians specialised in museology and the historians of IT who are responsible for the exhibits, who need to learn to understand one another's requirements so that the exhibition space is equipped and lit in order to throw into relief working machinery (not static objects as in other exhibitions).

Preservation for the future: augmented reality

An exhibition on simulation takes advantage of IT technologies, and in return it gives new options for developing these technologies.

VR about VR

Records of the exhibition will be preserved as 'augmented reality', a project which will maintain a permanent record of the exhibition. The architectural setting of the exhibition is suited to VR, so that the result will be the history of VR within a VR model. A narrative structure will have to be combined with a virtual spatial environment. The project is thus a contribution to theories on the structure of digital information.

Records for preservation

While the exhibition was on, records were made: photos, films of the lectures, which will be integrated

with the documents shown, originally in several formats. The recording of the exhibition in progress for future VR use was part of the sessions themselves.

Visualisation thus combines aesthetic and technical issues at several levels. The poster will give images for each point; the several types of computer programmes presented with it will allow the conference delegates to experiment on demos of work in progress as well as see views of the exhibits at different scales.

References

- Bonnett, J.** (2004). 'New Technologies, New Formalisms for Historians: The 3D Virtual Buildings Project'. *Literary and Linguistic Computing*. **19(3)**: 273-287.
- Brito, A.** (2008). *Blender 3D: Architecture, Buildings and Scenery – Create Photorealistic Visualizations of Buildings, Interiors and Environmental Scenery*. Birmingham: Packt Publishing.
- Centre for Advanced Studies in Architecture, The University of Bath.** (2010). *Completed Projects, Current Research*. <http://www.bath.ac.uk/casa> (accessed 20 March 2010).
- Gallet-Blanchard, Liliane.** (2009). 'From Virtual Reality to Augmented Reality and 'Augmented Virtuality''. *Le Virtuel, une Réalité? Is the Virtual Real?*. http://www.feb-patrimoine.com/nsdat/mmediatheque/expos/sorbonne_2009/conferences/LGB-siteFEB.htm (accessed 20 March 2010).
- Houdin, J-P., Tran, F.** (2009). *Khéops Révélé*. Paris: Gédéon – Dassault-Systèmes, DVD.
- Supercomputer Challenges*. **408** June 2007, special issue.
- Tavernor, R., Day, A.** (2000). 'The Computer and Urban Modelling: Experiences Through Time'. *Villes en Visite Virtuelle*. Gallet-Blanchard, L., Martinet, M-M. (eds.). Paris: Presses de l'Université Paris-Sorbonne, pp. 21-20.
- Terras, M.** (2000). 'Virtual Reality and Archaeological Reconstruction'. *Digital Environments: Design, Heritage and Architecture, CHArt Conference Proceedings, vol.2..* <http://www.chart.ac.uk/chart1999/papers/noframes/terras.html> (accessed 20 March 2010).
- King's Visualisation Lab.** (2010). *Projects*. <http://www.kvl.cch.kcl.ac.uk/> (accessed 20 March 2010).

Discursive Metadata and Controlled Vocabularies

Mylonas, Elli

elli_mylonas@brown.edu
Brown University, USA

Wendts, Heidi

heidi_wendt@brown.edu
Brown University, USA

Bodel, John

john_bodel@brown.edu
Brown University, USA

While formulating an Epidoc compliant template for the encoding of ancient inscriptions, it became apparent that it was necessary to accommodate discursive information about various characteristics of an inscription as metadata in the header of a document, and to specify the same characteristics using a controlled vocabulary, to facilitate searching, sorting and indexing. The msDesc features of the TEI guidelines do not actually allow this type of encoding to occur in several crucial places. However, it is possible to achieve both goals by repurposing, and perhaps straining the usage of some TEI features. We will describe the problem and our solution in more detail, in order to document one project's solution to a common problem, but also to suggest that the TEI Guidelines might be modified to allow this as a more normal use.

Epidoc, a TEI P5 schema that has been developed for epigraphical and papyrological materials is widely used for encoding classical and other western inscriptions. Historically, there have been two parallel and converging ways to encode this type of documentary evidence. The first treats the transcription of the text together with descriptive information about the support, context, decoration and history as content, the way it might be if it were published in a book, and enclosing it all within the <text> element. In this type of encoding the TEI header information is brief, serving to document the source publication, and not the inscription. The primary example of this type of encoding is InsAph, which originated as the digital version of a print volume, and represented the publication of record for its inscriptions. The second approach treats the text of an inscription as content, and places contextual information such as the description of the surface the inscription was written on, its date, format and origin as structured metadata in the TEI header. US Epigraphy, which originated as an aggregation of inscriptions, most of which had already been

published, is an example of this approach. These approaches have different advantages: the first results in a more readable and more nuanced description of the inscription. The second, in which the placement of information is more predictable and controlled, allows better processing, searching and indexing.

The US Epigraphy project records Graeco-Roman inscriptions that are known to be in United States collections so that they may be located and studied or used in teaching. As such, the metadata that allows the inscription to be searched and sorted by its characteristics is of paramount importance. The project is developing its corpus using an iterative process, by which inscriptions are first recorded as an ID number with bibliographic citations, then metadata and images are added. Transcription and more detailed descriptions, a necessarily slower process requiring more epigraphical expertise, are added as a third step. This progression ensures that the corpus is as complete as possible, and that information is added in a sequence that provides as much information as possible about as large a number of inscriptions as possible.

US Epigraphy, following the TEI P5 version of the Epidoc schema and encoding practices, relies heavily on the TEI header and uses the `<msDesc>` component of the header to record metadata about an inscription.

In `<msDesc>` there are elements to indicate the genre to which the inscription text belongs (`<msItem class="xx">`), the type of support on which it is inscribed (`<objectDesc form="xx">`) and the material of which it is made (`<supportDesc material="xx">`). These three elements are used to indicate parallel types of information, but unfortunately, they don't exhibit parallel behaviors.

`<msItem>` has an attribute to indicate text genre and it can accommodate more discursive detail in a child `<p>`. The attribute, `@class`, is a specialized attribute of type "data.code" that allows the `msItem` to point to a controlled vocabulary of text genres. This is handled through a complex mechanism as follows: the text genres are listed using a `<taxonomy>` element in `<profileDesc>`. `<textClass>`, also part of `<profileDesc>` then points to a genre in the taxonomy, and `msItem/@class` in turn, points to `<textClass>`. This is complicated, but it allows a controlled, and less precise value to co-exist with a more nuanced but less processable description of the text genre. Also, crucially, it maintains the controlled list of genres in the document, and not in the schema.

Conversely, `<objectDesc>` and `<supportDesc>` have specialized attributes `@form` and `@material` whose values belong to the class "data.enumerated," forcing their values to be maintained in the schema. This

is undesirable, as it means that an encoder, or a content specialist would have to modify the schema in order to change a controlled vocabulary. Changing the values in an enumerated attribute also means that it will no longer be possible to validate different epigraphical projects with the same schema, even though their document structures are fundamentally the same.

The ideal solution is to be able to maintain taxonomies within the document, and refer to values within them using an attribute such as `@ana`, whose value belongs to class "data.pointer." `<taxonomy>` provides a powerful and appropriate classification structure, but in the guidelines it is defined as containing only information on text genres, and forming part of the `<msItem><textClass>` construct. `@ana` can point to interpretive elements such as `<interp>` and `<fs>`, but not `<category>`, which is the constituent part of `<taxonomy>`.

Currently, it is possible to create several taxonomies, and to access them using the `xi.include` mechanism, so that all files and all encoders are using the same controlled vocabularies at all times, and updates are immediate. It is also possible to point to elements in the taxonomies from `<objectDesc>` and `<supportDesc>` using an `@ana` attribute, since `@ana` is globally available, and points to a valid URI. However, although this validates, it isn't semantically correct according to the TEI guidelines. A more satisfying solution is to redefine specialized attributes like `@support` and `@material` to behave like `@ana`, and be able to point to controlled vocabularies such as those contained within `<taxonomy>`.

It is important, when encoding highly structured but also potentially idiosyncratic materials, like inscriptions or papyri, to be able to use both controlled and full-text descriptions. This should be enabled by the markup, but should also be encouraged as good encoding practice. It is also expedient and easier to avoid errors for encoders and programmers to have parallel structures describing similar types of information.

As corpora like US Epigraphy, InsAph, DDDP and similar collections become more concerned with how they will be mined and processed, and are no longer content with creating digital facsimiles to facilitate access, this type of information management is becoming more important. This poster has focused on a few elements and their accompanying attributes. They are not the only places where this problem arises, however. The solution that is presented here is by no means an ideal one. Indeed, it is only permissible insofar as it results in valid TEI documents. There are several other possible approaches. The best solution will be one that results

in a set of best practices that can be re-used in other, similar situations.

References

- Epidoc.* <http://epidoc.sourceforge.net/> (accessed 11/2/2009).
- InsAph.* <http://www.insaph.kcl.ac.uk/index.html> (accessed 11/2/2009).
- TEI Guidelines.* <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html> (accessed 11/2/2009).
- (11/2/2009). *US Epigraphy.* <http://usepigraphy.brown.edu>.

The Digital Ark: From Taxonomy to Ontology in 17th-century Collections of Curiosities

Brent Nelson

brent.nelson@usask.ca
University of Saskatchewan

When the famous seventeenth-century gardener John Tradescant named his home, with its collection of rarities and curiosities, “the Ark,” he was expressing his desire to compile a microcosm of a wide world of variety beyond common experience. Such collections represented the sum of early modern European experience of the world at a time of rapid scientific and geographical expansion and reflected fundamental epistemological shifts in attitudes toward curiosity, wonder, and credulity on the cusp of the modern age. The rapidly expanding world of exploration, colonization, and commerce in the seventeenth century proliferated with strange and bizarre creatures and artifacts that challenged the traditional limits of knowledge. To meet the need for a complete and accessible record of early modern collections, ‘The Digital “Ark”’ will accumulate a database of artifacts and natural specimens as represented by documentary records of early modern collections (inventories, diaries, correspondence, etc.), contemporary drawings and engravings, as well as digital images and curatorial records of extant remnants of these collections. It will be an extensive record of all known collections of rarities and curiosities in England and Scotland from 1580-1700 for which documentary evidence survives, comprising up to 10,000 specimens and artifacts. This information, both textual and visual, will be delivered in an open-access Web-based virtual museum that will collect and display artifacts and natural specimens drawing from a fully searchable database that will record and classify these items and their descriptions in some two dozen fields of information.

This poster will briefly introduce the project and then focus on the challenges this data poses for a computational process that involves naming data types and defining relationships between them. The principle challenge comes from two unique aspects of the project:

1. The need to accommodate in the user interface a wide range of source genres in a rationalized and consistent form, while representing the distinct

- epistemological modes of these diverse forms of representation;
2. The need to respect and reflect the way the data was viewed and understood in this age of transition between humanistic and empirical ways of knowing as we interpret the data set and design a database structure to encode, store, and represent this data in all of its complex relationships.
- The poster will have four sections:
1. An introduction providing a brief paragraph on the cultural background illustrated with a bulleted set of statistics and a 17th-century engraving of a typical cabinet of curiosities.
 2. A chart depicting the diverse data types that provide the content of the digital ark, including: letters; travel accounts; diaries, inventories and catalogues; discursive prose; poetry; contemporary engravings; drawings and paintings; modern photographs of extant objects; and secondary scholarly sources) along with the characteristics of these genres that complicates the process of defining data structures. Page facsimiles will illustrate these data types.
 3. A chart depicting the differences between a taxonomic and an ontological view of data. In brief, the taxonomic approach involves entry into a new body of data and the naming and categorizing process that occurs as one interprets and makes sense of this new data, while the ontological approach involves fixing categories and properties in a determined order of being. The seventeenth century represented a significant shift from ontology, where the nature of existence was received and commonly understood by all, to an age of taxonomy, where the new and strange demanded an open-ended reconsideration of the world of existence and a continual configuration of knowledge. In the computer age, we are experiencing a similar tension in the desire to explore and discover relationships between data, while at the same time thinking of data representation in terms of ontologies. This chart will represent this tension both in the context of the epistemology of the early modern collections and the context of computer processes that might be employed to represent them.
 4. The conclusion will outline two steps that will be taken to address these needs:
 - i. The use of qualitative tagging as a means to interrogating the source documents to find what is there, before determining the tag set that will inform the final data structure, that is, to infer a taxonomy rather than simply impose an assumed ontology.
 - ii. The use of a combination of the TEI structure to represent text-based sources with modified object-based ontologies to represent the objects as depicted in these textual sources and also as depicted in graphical sources, both contemporary engravings and drawings, and modern photographs of extant objects.

Acknowledgments

This project is funded by the Social Sciences and Humanities Council of Canada.

References

- Bradley, John** (2005). 'Documents and Data: Modelling Materials for Humanities Research in XML and Relational Databases'. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing*. **20.1**: 133-51.
- Corns, Thomas N.** (2000). 'The Early Modern Search Engine: Indices, Title Pages, Marginalia and Contents'. *The Renaissance Computer: Knowledge Technology in the First Age of Print*. Neil Rhodes, Jonathan Sawday (eds.). London, England: Routledge, pp. 95-105.
- Daston, Lorraine** (2007). *Objectivity*. Peter Galison (ed.). New York: Zone Books.
- Daston, Lorraine** (1989). 'The Museum: Its Classical Etymology and Renaissance Genealogy'. *Journal of the History of Collections*. **1.1**: 59-78.
- Gilbert, Neal Ward** (1960). *Renaissance Concepts of Method*. New York: Columbia University Press.
- Grindle, Nick** (2005). "No Other Sign Or Note than the very Order": Francis Willughby, John Ray and the Importance of Collecting Pictures'. *Journal of the History of Collections*. **17**: 15-22.
- Harmon, Margaret** (1975). *Stretching Man's Mind: A History of Data Processing*. New York: Mason/Charter.
- Leith, Philip** (1991). 'Postmedieval Information Processing and Contemporary Computer Science'. *Media, Consciousness, and Culture: Explorations of Walter Ong's Thought*. Bruce E. Gronbeck, Thomas J. Farrell, Paul A. Soukup (eds.). Newbury Park, CA: Sage Publications, pp. 160-176.
- Liu, Yin, Jeff Smith** (2008). 'A Relational Database Model for Text Encoding'. *Rept. Digital Studies/Le champ numérique*. **0.11**.
- MacGregor, Arthur** (c2007). *Curiosity and Enlightenment: Collectors and Collections from the Sixteenth to the Nineteenth Century*. New Haven, CT: Yale University Press.

Marcus, Leah (2000). 'The Silence of the Archive and the Noise of Cyberspace'. *The Renaissance Computer: Knowledge Technology in the First Age of Print*. Neil Rhodes, Jonathan Sawday (eds.). London, England: Routledge, pp. 18-28.

Obrst, Leo, Howard Liu (c2003). 'Knowledge Representation, Ontological Engineering, and Topic Maps'. *XML Topic Maps: Creating and using Topic Maps for the Web*. Sam Hunting (ed.). Boston: Addison-Wesley, pp. 103-149.

Sowa, John F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove: Brooks/Cole.

Stafford, Barbara Maria (1996). *Good Looking: Essays on the Virtue of Images*. Cambridge, Ma.: MIT Press.

"Inventing the Map:" from 19th-century Pedagogical Practice to 21st-century Geospatial Scholarship

Nowviskie, Bethany

bethany@virginia.edu

Scholars' Lab, University of Virginia Library

In 1823, at a small school in western Vermont, Frances Alsop Henshaw, the 14-year-old daughter of a prosperous merchant, produced a remarkable cartographic and textual artifact. Henshaw's "Book of Penmanship Executed at the Middlebury Female Academy" is a slim volume, later bound in marble boards, containing – in addition to the expected, set copy-texts of a practice-book – a series of hand-drawn, delicately-colored maps of our nineteen United States, each one paired with an edited, geometrically-designed and embellished prose passage selected from the geography books available to a schoolgirl in the new American republic.¹ Henshaw's maps and texts alike are interpretive re-presentations of this body of contemporaneous geodetic and descriptive literature. Formally, many of the textual passages that accompany her maps are designed within a framework of aesthetically-inflected cardinal coordinates, representing (either conceptually or in their spatial contours) the states they describe, and positioning political and natural boundaries in cartographically appropriate margins of the page [see Figures 1 and 2].

As a work of juvenilia, Henshaw's "Book of Penmanship" is no less remarkable in its artistic and imaginative accomplishment for being exemplary of larger trends in the geographic education of nineteenth-century Americans. A sampler in codex form, the book constitutes a set of interrelated pedagogical and personal exercises in geospatial and textual graphesis, or subjective knowledge-production through the creation of images and texts-as-image. Drawing exercises of this sort were developed by noted American educator Emma Willard, founder of Henshaw's Vermont school and author of several geography textbooks. In a period when reading and recitation of geodetic texts were presumed the best aids to spatial memory, Emma Willard believed that students should learn through the personal creation and analysis of drawings – even, or perhaps especially, of drawings that embed subjectivity and aesthetic choice. ("In history," wrote Willard, with characteristic confidence, "I have invented the map.")² My presentation argues that

attention to the processes and products of Willard's pedagogy can be as fruitful for modern scholars, who grapple with the integration of geospatial technologies into the interpretive humanities, as geographers and literary historians demonstrate them to have been for meaning-making among an increasingly spatially-literate populace in the early years of the American republic (See Brückner, 2006; Schulten, 2007).

Work and interest in the geospatial humanities is growing – at a variety of scales, and with a variety of institutional inflections – in libraries, academic departments, and digital centers around the world. Despite the richness of this activity, scholars press up against a well-documented series of obstacles, pragmatic and conceptual, in their use of spatial tools, datasets, and methods.³ In the ongoing interchange of the digital humanities, could new methods and self-consciously literary and ludic perspectives permit us, with Emma Willard, to *invent the map?*

An examination of spatial decision-making and of the interplay among text, image, and geographical source material in the Henshaw book may suggest relations among her enterprise and some ambitions held by modern humanities scholars for geospatial technology. These relations hinge on an openness to graphesis and iterative design as a legitimate method in digital scholarship. I will also argue that a fresh, steady look at cartographic and geospatial technologies for the digital humanities should not be taken alone in the context of spatially-oriented disciplines (such as anthropology, area studies, archaeology, and environmental history) that have more traditionally made use of these tools and datasets and have, to greater and lesser extents, made peace with their present limitations – a set of assumptions that underlie and circumscribe the analytical and expressive power of geospatial information systems (GIS). Instead, I want to extend our examination of GIS technologies and the administrative, pedagogical, and scholarly publishing systems that support them *into the realm of interpretive literary and textual studies* – and imagine them at a variety of scales: from support for a complex mapping of print-culture production and distribution networks through space and time; to the visualization of subjective spatial expression in historical and literary documents; to an examination of the spatial and typographical features of a single page, or class of page designs. What potential might geographical tools and methods have for illuminating the spatial, semantic, and intertextual features of books as well as landscapes? Can we imagine a next generation of these tools in support of visual and aesthetic methodologies for very traditional (and, in

some cases, only marginally geospatial) humanities interpretation?

If our aim is to promote, among colleagues in fields like literary studies and digital history, a new and timely engagement with geospatial visualization as *interpretive practice* (timely both in terms of the burgeoning development and use of what have been called "vernacular" or crowd-sourced spatial datasets and interfaces outside of the academy, and in the context of a growing interest in a return to pragmatic, methodological training in graduate education within it),⁴ we must ask the following question: what is required of our shared tools, methods, and pedagogical practices to allow us to make as meaningful a visual intervention in our current scene as Emma Willard did in hers?

The deficiencies of existing geospatial applications and the social and academic systems that support and promote their use have been adequately surveyed. Martyn Jessop provides a thorough summary in the pages of *LLC*, when he identifies four factors contributing to a strange "inhibition" of the use of geospatial information among digital humanists, a community not generally daunted by the need to learn new software tools, metadata standards, and data curation practices (Jessop, M., 2008). The "first and most fundamental" of these inhibiting factors "concerns the use of data visualization and images *per se* in the discourse-based research methodology of the humanities" (42). That most humanities disciplines only make superficial use of images and image-based methodologies suggests an opportunity, if not a need, to interrogate our habitual interpretive practices and the ways in which graduate education perpetuates a longstanding marginalization of the visual – particularly infelicitous in light of the opportunities for production and analysis afforded by new media. Other factors involve: the suitability of current geospatial software packages to the treatment of issues like subjectivity and emotion, temporality as experienced and expressed in the documentary record, or interpretive inflection in the humanities; and those specific qualities of humanities information unsuited to tools that have been designed for synchronic analysis of incredibly dense datasets (rather than for sparse, temporally-inflected data) and with a scientific eye toward filtering out – rather than celebrating and analyzing – uncertainties or ambiguities. Finally, Jessop treats broader issues of scholarly communication: issues in funding, producing, evaluating, and distributing innovative geospatial scholarship in disciplines whose structures evolved in response to different conditions and expectations. With Jessop, I will suggest that, "although we usually think of GIS as a positivist tool its greatest contribution to the humanities... may be not as an analytical or

information presentation tool but as a reflexive one," allowing us not only to engage with the "highly experiential" and qualitative features of our datasets, but also to reflect on how we construct our disciplines (46).

Frances Henshaw's "Book of Penmanship" – a sophisticated, if naïve, 1820s pen-and-ink GIS – serves here as an example of both an illuminative process for, and a potential exemplar product of, a potential hermeneutic involvement on the part of scholars with textual surrogates and geospatial interfaces. We lack digital tools expressly crafted to promote the kind of ludic, iterative, graphical engagement with book design and geographical expression that is everywhere evident in the Henshaw artifact. But the components of these tools are all around us. It is less a technical than an institutional and intellectual problem to identify the small pieces – and practices – that must be loosely joined in order for humanities scholars to move forward in the arena of geographic and textual graphesis, or knowledge-making through graphical expression (Drucker, 2001; 2009).

Is there a methodological approach that presents itself as a way to crack open analytically – or perhaps just allow us to *replicate* and *play* in digital environments with – the easy brand of spatial and literary intertextuality evinced in Henshaw's schoolgirl exercise? I will look at a several classes of tools and digital humanities practices as a way of getting at this question, including: the iterative, interpretive, and structured sketching prototyped in Temporal Modelling (Drucker, J.D. and Nowviskie, B., 2004) and Neatline;⁵ data-mining for geography in massive text corpora through tools like MONK and TAPoR, and what the Google Books research repositories and efforts like HATHItrust must enable in their APIs to contribute to this field;⁶ textual and graphical collation interfaces predicated on visualization rather than – or as much as – on structured markup, such as Juxta and Sappheos;⁷ mobile, GPS-powered tools and toys; and powerful, analytical GIS applications like the ESRI products, not at all designed for textual studies, but ready nonetheless for some dedicated gate-crashing.



Fig. 1: Connecticut, one of 19 maps in Frances Henshaw's "Book of Penmanship Executed at the Middlebury Female Academy," 29 April 1823. Library of David Rumsey.



Fig. 2: Descriptive and positional text accompanying the Connecticut map; Frances Henshaw, 1823. Library of David Rumsey..

References

- Brückner, Martin.** (2006). *The Geographic Revolution in Early America: Maps, Literacy, and National Identity*. University of North Carolina Press.
- Drucker, Johanna** (2009). *SpecLab: Digital Aesthetics and Projects in Speculative Computing*. MIT Press.

Drucker, J.D. and Nowviskie, B. (2004). 'Speculative Computing: Temporal Modelling'. *A Companion to Digital Humanities*. Susan Schreibman and Ray Siemens (ed.). Oxford: Blackwell, pp. 431-447.

Drucker, Johanna (2001). 'Digital Ontologies: The Ideality of Form in/and Code Storage—or—Can Graphesis Challenge Mathesis?'. *Leonardo*. **34:2**.

Jessop, Martyn (2008). 'The Inhibition of Geospatial Information in the Digital Humanities'. *Literary and Linguistic Computing*. **23:1**.

Ramsay, Stephen (2005). 'In Praise of Pattern'. *TEXT Technology: the journal of computer text processing*. **Volume 14, Number 2**.

Schulten, Susan (2007). 'Emma Willard and the graphic foundations of American history'. *Journal of Historical Geography*. **33**: 542-564.

Notes

1. Accession #2501 in the Rumsey collection: <http://davidrumsey.com/>
2. Emma Willard to "Miss Foster," 5 November 1848. Reprinted in Lord's Life of Emma Willard. New York: 1873; page 228.
3. See the executive summary and report of the 7th annual Scholarly Communication Institute, Charlottesville, VA, 2009. SCI 7 focused on "Spatial Technologies and the Humanities": <http://uvasci.org/> and <http://uvasci.digress.it/>
4. "Vernacular" is perhaps a poor term to address commercial, cloud-based, and "neo-geo" tools and interfaces based on mobile technologies, GPS, virtual globes, and Web-based slippy maps. See Scholarly Communication Institute 7 report on "vernacular" technologies & reports from SCI 6 and 7 on methodological training: <http://uvasci.org/>
5. Neatline is a tool for the creation of interlinked timelines and maps as interpretive expressions of the literary or historical content of archival collections, currently under development by the Scholars' Lab at the University of Virginia with generous funding by the NEH: <http://neatline.org/>
6. In this area of activity, see Stephen Ramsay (2005) on "computational analysis in literary studies as a quest for interpretations inspired by pattern," which can move the "hermeneutical justification of the activity away from the denotative realm of science and toward the more broadly rhetorical and exegetical practices of the humanities."
7. See <http://juxtasoftware.org> and <http://sapheos.org/>

An Open Source Toolkit for Flexible Browsing of Historical Maps on the Web

Ohno, Shin

ohnoshin@gmail.com
Ritsumeikan University

Saito, Shinya

Ritsumeikan University

Inaba, Mitsuyuki

Ritsumeikan University

Map browsing on the web is now commonly used, and most people have used map services such as Google Maps, Google Earth, and Yahoo! Maps. These services also provide developer API, which makes it easy to integrate map services into the website. Map browsing in practical use is covered with those services, but displaying historical maps as exhibits is not, which requires different approaches.

We, three members of Web Technology Research Group of the Digital Humanities Center of Japanese Arts and Cultures, Ritsumeikan University have developed an image viewer, which can handle zooming, panning, and rotating. Although there exist many similar toolkits, none has all three features with JavaScript written. Furthermore, we would like to share it as open source library.

This paper starts with elucidating Japanese historical maps, their unique features and needs. After that, it discusses our image viewer, its design and implementation to meet their needs.

1. Purpose

Since Japanese historical maps do not have concepts of geo-coding usually and are distorted, it is difficult to integrate them with map services such as Google Maps.

Using Keyhole Markup Language (KML) makes it possible to integrate these maps with Google Earth. There is some research on historical maps on Google Earth (Zeile et al. 2007Nishimura et al. 2007). Although those tools are highly functional for defining 3D geographical information on the web, the user interface is not easy to use for traditional humanities scholars who are normally not familiar with KML language. Moreover, setting the image in Google Earth is not always their priority. Therefore, when it is necessary to exhibit original, hand-

drawn maps on the web, one has to take different approaches to integrating these images into the web.

Because of these issues, image viewers are usually used to implement historical maps for exhibition. These viewers enable users to zoom images, and to see their details if they are provided in high quality. Moreover, today's high technology for user interface allows us to use such viewers with regular web browsers.

Since most of these viewers are not open source based applications, however, no room exists for us to extend and add more features. Even though they are provided as open source, most of them depend on a Flash plug-in, and there is no implementation with JavaScript. Because JavaScript is lightweight and easy to be integrated into websites, we have decided to implement this map viewer with JavaScript, and open it to the public as an open source product, and we believe there should be alternative options for us as open source based toolkits for these needs.

2. Design



Figure 1: Sample of Japanese historical map

In this section, we discuss what kinds of features are required for historical map viewers, and compare them with related software.

When displaying Japanese historical maps as image exhibits on the Web, we encounter difficulties in handling angles. With no clear concept of top or bottom in these maps, names of places are labeled from many directions. This is related to the fact that people in those days put rather a large map on the table or the floor to read it. Because of this particular feature of the historical maps, as opposed to maps today, browsing tools for these historical

maps should provide a way of viewing them from multiple angles. Therefore, we decided to design our tool to support flexible browsing rather than geographical accuracy.

What we found out about other image viewers is that many can support zooming and panning, but not rotating. Most of them also depend on a Flash plug-in, and there are no implementations with JavaScript. There are comparisons of similar software for viewer of maps or images.

- *Google Maps/Google Earth* Since their beginning, these services with web browsers have become more commonly used, and more people have become used to them, which in turn has changed the way we access geographical resources. However, historical maps do not have accurate longitude and latitude, and even using KML to integrate them into Google Earth, the images are distorted and we need to change the purpose to display them. Plus, KML is rather complicated to maneuver it very well.

	API	Free	Rotat-able	Zoom-able	Image Cutting	Open source
Google Maps	Yes	Yes	No	Yes	Yes	No
Google Image Cutter	No	Yes	No	Yes	Yes	Yes
Google Earth	Yes	Yes	Yes	Yes	Yes	No
Open Zoom	No	Yes	No	Yes	Yes	Yes
Open Layers	Yes	Yes	No	Yes	Yes	Yes
Ojikit Image Viewer	No	Yes	Yes	Yes	No	Yes

- *GMap Image Cutter by LCUGMap Image Cutter by LCU:* To use Google Maps image viewer, LCU developed GMap Image Cutter tool to cut a large image into many small pieces. With this Cutter provided, users can integrate images to their own website with the same Google Maps user interface.
- *OpenZoom:* OpenZoom is a free and open source toolkit for delivering high-resolution images and zoomable user interfaces. Like GMap Image Cutter, it can cut a large image into many small file-size images. With this software, users can download an appropriate file size to their local computer, which works to prevent network traffic from congesting in cases in which the image is too large to download. This is SWF based, and requires a Flash Plugin.
- *OpenLayers:* OpenLayers is a free and open source library for handling geographical information written in JavaScript. Since OpenLayers implements both WMS(Web Map Service) and

Web Feature Service, users can integrate many different maps service to their own website. It has zooming and panning features for browsing maps on the web.

- *Ojikit ImageViewer* Ojikit Image Viewer is the toolkit we have developed. It has zooming, panning, rotating features, as well as an open source license. While it does not have the function to cut a large image into smaller files, improvement of broadband allows us to transfer files without so much reduction. Moreover, one file is easier to store and manage than many files.

3. Implementation

Giving consideration to what historical maps viewers need to be, we have developed an open source toolkit. This toolkit is written in JavaScript, and it has lightweight features and an open source license. Written in JavaScript, Mozilla and Safari based browser uses SVG, and the Internet Explorer uses VML to handle vector images. To develop this library seamlessly in different browsers, this library depends on jQuery and Raphael.js.

4. Application

We have applied this tool to Japanese historical maps as our Center has a large number of historical map images, open to the public. Unlike other image viewers, which do not support rotating function, this toolkit enables us to rotate all images. Figure 2 shows one example of these map images as they look on the screen. With this toolkit, users can read labels written from many directions, as they are on the map.



Figure 2: The left map is rotated 130 degrees, and the right map is rotated 210 degrees.

5. Conclusion

Addressing a culture-specific problem of Japanese historical maps, this paper discusses current map viewer systems to deal with this problem. This toolkit was developed as a viewer for Japanese historical maps per se. However, its extensive and easy integration feature leads us to consider some other possibility to apply it for other fields that have similar demands such as rotation and zooming features with easy integration. While this toolkit's development is our on-going project, its source code is meanwhile open to any users.

There are three future areas of research related to our toolkit. First, we have applied our toolkit to Japanese old maps browsing as a test this time, and we are trying to produce more tests with the toolkit. Secondly, related to the first point, we think that this toolkit is useful not only for maps, but also for other images that need to be browsed from many different directions, such as handwriting manuscripts, and apply different resources. Finally, this toolkit does not consider geo-reference. In our center, there is a group for GIS research, and for future provision, we are discussing how to co-reference its geographical information with each other, so that this tool can represent geo-reference.

References

- Nishimura, Y. et al.** (2007). 'Analysis of Silk Road Old Maps Using Google Earth'. *SIG Computers and the Humanities Symposium 2007*. Information Processing Society of Japan, pp. 155- 162.

Zeile, P., Farnoudi F., Streich, B. (2007). 'Fascination google earth - use in urban and landscape design'. *Embodying Virtual Architecture: The Third International Conference of the Arab Society for Computer Aided Architectural Design (ASCAAD 2007)*. Pp. 141-148.

Google Maps. <http://code.google.com/intl/en/apis/maps/> (accessed 10 October, 2009).

Google Earth. <http://earth.google.com/intl/en/> (accessed 10 October, 2009).

GMap Image Cutter. <http://www.casa.ucl.ac.uk/software/googlemapimagecutter.asp> (accessed 10 October, 2009).

OpenZoom. <http://openzoom.org/> (accessed 10 October, 2009).

OpenLayers. <http://openlayers.org/> (accessed 4 March, 2010).

Text-Image linking of Japanese historical documents: Sharing and exchanging data by using text-embedded image file

Okamoto, Takaaki

04c0004@sch.otani.ac.jp

Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University

This paper will demonstrate the effectiveness of the linkage between textual data and image data in the field of medieval Japanese history where the research primarily uses text-based documents instead of photocopies and digital images.

1. Using Computers and Data for Historical Studies

The current state of the field of Japanese history requires that researchers (1) view more historical documents than in the past, (2) fully utilize the various types of information contained in these documents such as the form of a character, handwriting, type of paper etc., and (3) collaborate with others on inter-disciplinary projects instead of pursuing isolated projects in a single field. It is not possible to realize these goals through the individual efforts of researchers. A well-established digital environment becomes necessary. The most important and fundamental research tasks for a historian include: (1) looking up certain texts: what documents include these and where in the documents they can be found, and (2) examining the target documents. Given the nature of such research work, the utilization of computers is not as advanced as it should be. Without a fully computerized environment to enable easy access, the groundwork of identifying and reviewing documents would take considerable time and effort, and thus impede the advancement of research. If images and texts of documents are digitally connected, by searching keywords or sentences, the system will display the target texts directly as a part of the image, and users would also be able call up lists of entire images by inputting single characters

2. Semantic information and visual information of Historical Documents

In the field of history, most commonly circulated materials are published in traditional printed media such as reprints of texts. Historical documents that have not been reprinted are used only when they are significant for specific research projects. For historical studies, much of the emphasis is on the interpretation and analysis of primary texts, and reprints of primary material, in plain text form, are often used for such purposes.

Inevitably, certain information such as the original handwriting will be lost in the process of transforming the original into plain text. In addition, handwritten material published during the pre-modern period used numerous variants of Chinese characters (異體字) different from modern publishing standards. Many of the variations of old characters are not represented by computerized fonts and thus are often lost in the conversion into text files. Kunten (訓點), the punctuation marks to read Chinese text (Chinese classics, sutras translated to Chinese, etc.) in Japanese pronunciation and grammatical order, are represented by special symbols. For research using this type of document, simple text files are not enough. The primary material needs to be converted into image files.

In other words, historical documents contain semantic information that can be converted into text files as well as visual information that cannot be represented in text files.

3. Possibilities of Visual information

To use handwriting data as an example, historical documents often do not indicate the author's name. Even in cases when a signature is included, it can often be the person who authorized the document and not necessarily the actual penman. For this reason handwriting becomes an important piece of information to determine the composer of the document. Graphology analysis can reveal answers to the following questions: (1) who manually composed this document, (2) why did this person write this document, and (3) what other documents have possibly been written by the same person? In conjunction with content analysis (interpretation), these additional aspects can further advance historical and diplomatic research. Introducing imagery analysis to research that has been centered on textual analysis will no doubt lead to new developments.

4. Computer Environment for Handling Historical Documents

Although researchers have long recognized the importance of visual information, limited research has been done in these directions. One main reason for this is the lack of a well-developed digital environment, which has made such work highly labor intensive.

In order to make searching digitized images just as convenient as searching digitized text, images need to be organized based on the questions of "what character is contained in what part of which document." It will also be necessary to provide easy to understand results that will highlight searching characters or text within the image. To enable such functions, text data needs to be correlated with image data through coordinates by each character. In other words, three types of input are needed to set up this coordinate system: (1) Which document text does this image correspond to? (2) What characters or words are included in the text? (3) Where in the image are the search words? Two of these inputs, the image data and the text data, are already in common use. If every individual character of the text can be linked to the image through a coordinate system, it will enable text searches within the image of a document.

5. Text-image Management Tool and File Sharing/Exchange via Portable Devices

Researchers need to build an easily accessible digital environment with the documents and images of the documents and would thus find it useful to have a tool that could indicate the position in the original documents for a given text of interest. We are currently developing a tool that processes image files of text documents in this way (Fig. 1). Using this tool, a researcher, by clicking on any character in the image, can create information about the character such as its position within the image. As shown in figure 2, the data created for each character including ID (automatically generated GUID), position in the text, coordinates, dimensions and so on are made into a simple text file that can be managed by external software applications like Excel.



Figure 1

chineserichter_gold	chineserichter_widmungen_Paragraphen_IndexIndex_layer	X	Y	Width	Height
e0d0a58a-5112-4714-9aa7-wa0cc0f937a0	東	1	8402	1307	850
e7234f11-0408-4f0b-a1e6-9ef0a76e4791	司	2	8402	1795	850
e91180a0-2e03-4151-909a-85a0ea17d74	被	3	8402	2162	850
e754a100-2649-4466-a274-22979667d7a	大	4	8402	2493	850
e994d095-e709-4e22-9f11-444705b99dc2	奇	5	8402	2719	850
e17f6295-ecce-45d1-9a73-3971e9f99	頌	6	8402	3401	850
e0a0ff1c-3282-46bc-9842-3e7140e7593	大	7	8402	3718	850
e30e0400-1111-4d1e-a5d2-051463b987	奇	8	8402	3952	850
e6a20a0e-533b-4a13-9a03-22358449	頌	9	8402	4297	850
e6ca87c2-7140-4146-922a-872709e549	東	10	8402	4604	850
e36393ba-e356-4a70-a74-a11201f900	司	11	8402	4874	850
e71032a1-e114-408a-a20f-1753909a294	被	12	8402	5165	850
e51089e4-5165-4105-936754a010e9547	大	13	8402	5504	850
e7a40239-700e-405b-9c23-74a5d44b5	奇	14	8402	5833	850
e5c45620-3c51-4a64-aabb-2792140e742	頌	15	8402	6142	850
e50a9782-e517-4162-9344-e297330014b	東	16	8402	6437	850
e3405559a-5379-4a36-9275956a1970	司	17	8402	6729	850
e1945a7-9344-411a-9a42-493a4939	被	18	8402	6973	850
e94295a7-773c-4818-9c86-7a4770533ea	大	19	8402	7168	850
e10444f1-c305-4f5a-9f72-89517a1fb9a	奇	20	8402	7362	850
e7446410-771a-4250-984c-891a74649f	頌	21	8402	7553	850
e4646783-1500-4f6e-971a-2255a744000	東	22	8402	7773	850
e0225790-3011-4520-9520-867181	司	23	8402	8014	850
e100877b-7-dab7-4a12-9221-2330388119c	被	24	8402	8234	850
e6bf4a02-a6d2-453c-9823-3a16d4c4	大	25	8402	8455	850
e6910009-940b-453c-9823-3a16d4c4	奇	26	8402	8649	850
e100877b-7-dab7-4a12-9221-2330388119c	頌	27	8402	8837	850
e00378a1-710c-47a6-9a0d-f7a95c02409	東	28	8402	9051	850

Figure 2

The program can generate reduced images for viewing in a browser and an HTML file containing information of the positions of the characters. When users search for a string of characters, a Javascript function will highlight the search results on the generated HTML file for display (Fig. 3).



Figure 3

The process of connecting these two sets of information cannot be automated and is dependent on manual input, which will inflict certain costs. Moreover, since the productivity of data processing by individuals is limited, group collaboration is essential. For the sake of simpler file sharing, this software thus creates one single file in TIFF format that combines the image data and the text data files. This enables the sharing of text data and image files through the exchange of just one file, an image file

embedded with text data through a portable device. Moreover, on the receivers' end, it will be possible for all the image files of the individual characters and the HTML files needed for searching text strings to be generated from the one text-data embedded image file, eventually making all of these exchanges unnecessary.

6. Publishing on the Web

Although owners of the documents are often unwilling to publish the images of their documents on the Web, it is still necessary to develop a network friendly environment for those documents to be published. Such an environment consists of web servers, databases, and web applications that connect them. Since the system needs to be user-friendly for the researcher, there is a tool for importing image files with embedded textual data as mentioned previously by a simple drag-and-drop function. Once the file is dragged and dropped into the tool, it will generate the following: (1) an image for display on the Web (2) coordinate axes for this image (3) partitioning of the image for zooming and update temporary text/index data in the database server. Then the image file and textual data will be ready for publication on the web.

7. Conclusion

This method uses data collected and compiled by individual researchers on their personal computers. Data sharing is done through external portable devices. Data sharing via external portable devices may cause delays in distributing the data, prevent other collaborators from referring to updated information, and ultimately cause inconsistencies in data references. The inconvenience of external portable devices may lead to situations where the data are not updated at all on the computers of other collaborators. To avoid these situations it is important to take advantage of the online environment. However, sometimes owners do not want researchers to put images of their historical documents in the environment. This paper demonstrates how to deal with such situations.

Knowledge and Conservation - Creating the Digital Library of New Hispanic Thought

Priani, Ernesto

epriani@gmail.com

Universidad Nacional Autónoma de México - UNAM, Mexico

Galina, Isabel

igalina@unam.mx

Universidad Nacional Autónoma de México - UNAM, Mexico

Martínez, Alí

Universidad Nacional Autónoma de México - UNAM, Mexico

Chávez, Guillermo

Universidad Nacional Autónoma de México - UNAM, Mexico

This paper presents the Digital Library of New Hispanic Thought (Biblioteca Digital del Pensamiento Novohispano). In Mexico, during the seventeenth century, two comets were observed, leading to the publication of various texts on the meaning of this astrological phenomenon (Trabulse 1994a). During this period relatively few books were published due to high costs, the required approval of the Inquisition for publication and the lack of potential buyers. This proliferation of publications on the subject can be viewed as an indication of the importance of these events for the Catholic Church, the colonial government and the intellectuals of the period (Trabulse 1994b). The Heavens were, in Spanish colonial Mexico, a scientific, religious, social, political and ideological battleground (Moreno 1998).

To this date most digitization projects for New Hispanic publications, namely Google Books and Biblioteca Virtual Cervantes, have focused primarily on access and preservation and although useful, do not constitute digital scholarly editions. The original aim of this project was to provide the basis for a critical digital edition through collaborative commentary of digital diplomatic transcriptions. The process however, in particular during the migration of our tailor made DTD to using TEI, also served as a research tool which enabled us to develop a conceptual framework to formulate further research on these texts.

This is the first project from our university to use XML and TEI.

1. From single publication to digital library

The original aim of the project was to make available in digital format the first book by Fray Diego Rodríguez, published in 1652. However, this idea was quickly superseded as we realized that, taking into consideration the relatively large number of books on comets and astrology published in this period, it was more desirable to present all texts as a corpus. We therefore decided to focus on the common theme of astrology/astronomy with the main objective of not only gathering and creating a corpora and providing access but also developing and publishing a digital critical edition with tools to aid scholars in their research. The outcome is the publication of all known texts of the period, eight in total.¹ With the exception of one, none of these texts has ever been reedited.

Up to now, due to the fact that the ideas contained within these texts were no longer of any scientific value, the text itself has not been considered particularly important. Moreover, of the few studies that exist on these publications, the tendency is to view them as general testimonies rather than concentrating on the texts themselves (Cf. Navarro, 1959 and Cullen, 1984). This new edition of the texts in digital format gave us the opportunity to examine the text in detail and reconsider previous interpretations.

2. Textual processing

2.1. Digitization of texts

The eight texts were transcribed into plain text digital format and diplomatic transcriptions were produced. The only text that had been reedited previously had been updated for spelling and grammar as well as correcting original errors, in order to make it more readable for a modern, non-academic audience (Cf. Navarro 1959 and Cullen 1984). Other editions had been facsimiles (Trabulse 2001). As the BDPN is aimed mainly at specialists and we wished to both preserve the texts as faithfully as possible and to provide analytical tools, diplomatic transcriptions were considered the most adequate.

2.2. Creating digital texts - DTD

The texts were then marked up in XML. We wished to approach the text with no preexisting theory or structure –and as a part of pedagogical approach to the use of XML as a research tool-, allowing the text to speak for itself. In order to do this we set

about producing categories as the text was analysed. These categories were then converted into tags which eventually formed the DTD for the XML mark up. The objective of this was that the analysis of the text for creating the DTD was in itself a critical analysis. Once our own DTD had been developed, we then began the process to migrate to TEI. In order to do so we sorted through our categories and found equivalents. This process helped us define a new conceptual framework for studying the texts.

3. Digital Library Tools

The main objectives of the BDPN web interface² were two fold: to present the marked up XML texts as a new digital edition and to offer tools to continue to carry out analytical work. Currently the system allows the user to view the published texts and consult them through indexes which are automatically generated from the XML tags (see Fig.1). An advanced search function allows the user to search specifically in over one hundred tags (see Fig.2)



Fig. 1 Index search interface



Fig. 2 Specific tag search interface

Additionally the system includes expert commentary on the text (see Fig.3). Specialists in the subject may obtain a username and password that allows them to comment on the texts and these appear as part of the critical edition. This provides the framework for future national and international collaboration on these particular texts. The project is not conceived as a closed edition and it will hopefully continue to be expanded to permit new enquiries of research to be developed.



Fig. 3 Screenshot of expert commentary

4. Discussion and Conclusions

4.1. DTD categories to TEI

The original DTD was used to identify all possible elements that could be of interest to the study; from classical mythology and Christian metaphysics to geographic and geospatial indicators. Almost one hundred categories were identified. Once the text was marked up using our DTD we then migrated to TEI Lite. In order to do so we needed to find equivalents for our elements. This migration process obliged us to refine our categories and interestingly to simplify our analytic framework. As we started using common TEI tags such as date, place, name, with the type attribute, we realized that we could eliminate many of our tailor made categories as they were redundant or too specific.

This process helped us to identify the truly relevant categories for these texts. For example, source (citation, commentary, authors), geographical and geospatial terms, diseases associated with comets and old Spanish terms. Significantly, the migration to TEI allowed us to identify all the comets mentioned in the corpus using the element attribution date. Now that we have identified the thematic structure of these texts using these categories, we now know that we wish to focus our research on astrological themes used, diseases commonly associated with comets, geographical areas affected by astrological predictions and the type of Spanish used in the period.

4.2. Documentation and technical development

A key issue in the development of this project was to produce enough documentation so that this can survive as an open ended project, allowing new texts to be incorporated when and if they are discovered. It is known many Humanities projects do not develop adequate documentation that allows text files to be reused (Warwick, et al. 2009).

Our experience shows that creating adequate documentation is extremely difficult. One of the main problems was the frequent changes in the tags. As the text was analysed and marked up, this tended to produce new requirements. It proved difficult to maintain the documentation up to date in order to describe and explain all the considerations and changes. This is a very time consuming process. It is difficult to anticipate these changes and to contemplate them adequately in the project timeline.

5. Conclusion and future work

The work done thus far has initially helped us to view all these texts as a network. This had not been done previously. Additionally in order to develop the tags we approached the texts with no preconceived structure and this has allowed completely new concepts to emerge from the texts. The move to TEI helped us to identify with precision the framework with which to approach further research into the text and produce a new reading of this corpus.

In the future we hope to incorporate the facsimiles as images and mark them up so that specific parts of the image can be associated with the textual transcription.

We also hope to generate international interest in these documents that up to now have been hidden in specialized libraries, mainly in Mexico. There is currently little critical digital content in Spanish in general and of this period in particular. We aim ultimately to generate new research on the subject and we believe that this tool can contribute to the study of New Hispanic thought. Additionally we aim to generate research and publications that also reflect on the experience of producing pre-modern electronic textual editions.

6. Annex 1- Digitized works

Escobar Salemerón y Castro, José de (1681). *Discurso Cometológico y relación del nuevo cometa*. México: Viuda de Bernardo Calderón.

Evelino, Gaspar Juan (1682). *Especulación astrologica y physica de la naturaleza de los cometas, y juicio del que este año de 1682 se ve por todo el mundo*. México: Viuda de Bernardo Calderón.

Kino, Eusebio Francisco (1681). *Exposicion astronomica de el cometa*. México: Francisco Rodríguez Luprecio.

López Bonilla, Gabriel (1675). *Discurso y relación Cometographica*. México: Viuda de Bernardo Calderón.

Rodríguez, Diego (1652). *Discurso etheorologico del nuevo cometa visto en aqueste hemisferio mexicano y generalmente en todo el mundo*. México: Viuda de Bernardo Calderón.

Rodríguez, Diego. *Modo de calcular cualquier eclipse de Sol y Luna según las tablas arriba puestas del mobimiento de Sol y Luna segun Tychon*. Manuscrito inédito.

Rodríguez, Diego. *Tratado General de Reloxes de Sol*. Manuscrito inédito.

Ruiz, Juan (1653). *Discurso hecho sobre la significación de impresiones meteorológicas que se vieron el año pasado de 1652*. México: imprenta de Juan Ruiz.

Sigüenza y Góngora, Carlos de (1690). *Libra astronómica y philosophica*. México: Viuda de Bernardo Calderón.

References

Carlos de Sigüenza y Góngora (1984). *Seis obras*. Cullen Bryant, William (ed.). Caracas: Biblioteca Ayacucho.

Moreno Corral, Arturo (1998). *Historia de la astronomía en México*. Mexico: FCE.

Trabulse, Elías (1994a). *Historia de la ciencia en México*. México: FCE.

Trabulse, Elías (1994b). *Los orígenes de la ciencia moderna en México. 1630-1680*. México: FCE.

Carlos de Sigüenza y Góngora (2001). *Libra astronómica y philosophica*. Trabulse, Elías (ed.). México: Sociedad Mexicana de Bibliófilos.

Carlos de Sigüenza y Góngora (1959). *Libra astronómica y philosophica*. Navarro, Bernave (ed.). México: Universidad Nacional Autónoma de México.

Warwick, C., Galina, I., Rimmer, J., Terras, M., Blandford, A., Gow, J., Buchanan, G. (2009). 'Documentation and the users of digital resources in the humanities'. *Journal of Documentation*. **65(1)**: 33-57.

Notes

1. See References.

2. Biblioteca Digital del Pensamiento Novohispano (<http://www.dbpn.unam.mx>)

Digital Forensics, Textual Criticism, and the Born Digital Musical

Doug Reside

dougreside@gmail.com

Maryland Institute for Technology in the Humanities, University of Maryland, USA

When Jonathan Larson, author of the hit Broadway musical *RENT*, died in 1996 just before his work opened off-Broadway, he left behind about 180 floppy disks containing, among other things, drafts of the musical composed over a period of about six years. These disks, which were donated to the Library of Congress and are now held there, represent one of the earliest examples of a "hybrid archive" - a collection of both paper and inextricably digital artifacts. Along with a series of timestamped Microsoft Word 5.1 documents, the disks also preserve early and transitional versions of the music in MIDI and MOTU Performer format that could not easily be transferred to a more traditional medium without significant loss. In this poster I show some of what I found on these disks, what it reveals about the creative processes that shaped *RENT*, and, more generally, how the lessons learned in my experience might be applied by others working with hybrid archives.

The earliest file on the Library of Congress disks relating to *RENT* is a version of the music for the title song timestamped 1:37 p.m. on December 21, 1989 and created with the musical editing program Performer (now called Digital Performer). Accessing this file was not an easy matter. I had planned to create an image of the disks using the "dd" disk imaging command built into most versions of Linux, but, unfortunately, the disks were formatted in the 800K HFS disk format and could not be natively read by a "modern" floppy drive. I therefore used a live CD install of Ubuntu 5.02 running on a Powerbook G3 to create the image. Of course, if I had not had access to this Powerbook, things would have been slightly more complicated. I could, perhaps, have brought a desktop with a third party floppy controller card (such as the Catweasel PCI card manufactured by Individual Computers), but getting such a bulky machine to the Library of Congress and through the airport level security would have been difficult. The Powerbook was an indispensable tool and well worth an eBay purchase for those doing similar work. Once the disk image was created I made a second working copy, mounted it on Mac OS X (the current version of the operating system still supports disk images in

legacy formats) and used a modern version of Digital Performer to open the file.

Note, however, that this digital file is not the earliest draft of *RENT* in the Library of Congress collection. There is a paper copy of the script that was probably written in mid-1989 by Larson's collaborator Billy Aronson. The draft is an 11 page, typescript that appears to have been produced on Aronson's letter-quality NEC printer. The draft is labeled "pre-lyric" and, true to this label, contains no songs but does include some relatively lyrical language (especially by Mimi who has lines like: "I embroider sunsets onto pillowcases. Well, now you know..."). The second draft in the collection, again paper and probably produced by the same typewriter used to produce the pre-lyric draft is labeled "Boheme" and dated 9/22/89. It assigns sole responsibility for the book and lyrics to Billy Aronson and the music to Larson and was, again, likely typed by Aronson. Most of the songs in this draft did not make it to the final version of the show, however the draft does contain versions of the songs "Rent" and "I Should Tell You" and, in more or less the form it is now known, "Santa Fe" (indeed, the program notes for *RENT* always credit Aronson for his work on these songs).

Although the broad details of the narrative that begins to emerge from this archive are well known (Billy Aronson and Jonathan Larson decided to collaborate on the musical, Larson initially only as composer, and together wrote three songs before going their separate ways), the digital and paper artifacts together fill out the story with precise and fascinating detail. For instance, although Larson probably received a script from Aronson by September (based on the date on the first script in the collection) and by November at latest (interleaved into the second draft is a letter from Aronson to Larson dated 12/1 which Aronson begins with the words "Here's the new last chorus for SANTA FE that you asked for"), Larson did not commit any work on the show to disk until December 21. The letter from Aronson indicates that Larson was probably working on the show before then, but likely recording his work, if at all, to analog media (perhaps, as he certainly did in other cases, to a cassette tape). However, in order to transcribe the music to digital format Larson required technology he did not have at home. Another Word Document on the disks dated 1/31/90 and named "STUDIO COSTS" appears to have been a kind of invoice to Aronson. It lists three studio visits, one for 6 hours on December 21 to create "Music Trax for SANTA FE & RENT," one for 4.5 hours on January 16 to create "Music Trax for I SHOULD TELL YOU" and one for 7.5 hours on January 30 to "Record Vocals for ALL SONGS" and to create "Mix Trax for ALL SONGS." That Larson sought out a digital studio so early in

the creative process (and was willing to pay about \$300 per session at a time when his primary source of income was part-time work at a diner) suggests how important Larson saw digital technology for his creative process. To truly understand *RENT*, then, the scholar must understand the digital technologies and processes used to create it. The textual critic of *RENT* and other born digital musicals must therefore be skilled in recovering, reading, and analyzing digital artifacts - the processes I hope to demonstrate in this introduction to my work with *RENT*.

Literary Theory and Theatre Practice: A Comparative Study of Watching the Script and the Simulated Environment for Theatre

Roberts-Smith, Jennifer

j33rober@uwaterloo.ca

University of Waterloo, Canada

Dobson, Teresa M.

teresa.dobson@ubc.ca

University of British Columbia, Canada

Gabriele, Sandra

sandrag@yorku.ca

York University, Canada

Ruecker, Stan

sruecker@ualberta.ca

University of Alberta, Canada

Sinclair, Stéfan

sgs@mcmaster.ca

McMaster University, Canada

Bouchard, Matt

matt.bouchard@gmail.com

University of Alberta, Canada

DeSouza-Coelho, Shawn

shawnathanddc@hotmail.com

University of Waterloo, Canada

Kong, Annemarie

aakong@yorku.ca

University of York, Canada

Lam, David

david.the.monkey@gmail.com

University of Waterloo, Canada

Rodriguez, Omar

omar.rodrigueza@gmail.com

University of Alberta, Canada

Taylor, Karen

katay164@interchange.ubc.ca

University of British Columbia, Canada

This paper describes the results of our recent work on a 3D prototype called the Simulated Environment for Theatre (SET), which we undertook based on a growing realization that our earlier 2D prototype, called Watching the Script, reified some fundamental

biases that would render it less useful for stage directors than we had originally hoped. Having produced the two prototypes, our next step is to carry out a user study to compare their usefulness. We intend to present the results of this user study, combined with a discussion of the meaning of the two designs.

The design of Watching the Script features three different perspectives on the text of a play: a microtext column that gives an overview of the length of sections, combined with coloured lines to indicate each character; a reading pane; and a dynamic playback on a stylized stage, where characters move around and their speeches scroll out underneath them.

While it has a certain naive charm, the Watching the Script prototype has several fundamental features that make it less than optimal for theatre directors. First is the degree of stylization of the stage, which is restricted to a single shape and does not readily lend itself to customization, either through changing the basic stage design or through applying details in the form of a set. Related to this stylization is the overhead perspective, which makes it difficult to imagine the actual lines of sight of people sitting in the audience.

Next is the association between movement and speech. Watching the Script is driven by the XML of the play, which means that character movements are attached to speeches. However, in actual practice, characters are not restricted to moving only when they speak. This emphasis on the speech as the fundamental unit of the play is reinforced through the central role of text in the interface, with each speech occurring simultaneously in three different places. We argue elsewhere (Gabriele et al. 2009) that both of these features make sense from the perspective of English literature, where the central object of study is the text. However, in the staging of a play, we recognize that the text, while still important, is a less central concern. Directors take the text of a play as a starting point, routinely cutting lines, removing entire scenes or characters, and so on.

Next is the association between movement and speech. Watching the Script is driven by the XML of the play, which means that character movements are attached to speeches. However, in actual practice, characters are not restricted to moving only when they speak. This emphasis on the speech as the fundamental unit of the play is reinforced through the central role of text in the interface, with each speech occurring simultaneously in three different places. We argue elsewhere (Gabriele et al. 2009) that both of these features make sense from the perspective of English literature, where the central object of study is the text. However, in the staging of a play, we

recognize that the text, while still important, is a less central concern. Directors take the text of a play as a starting point, routinely cutting lines, removing entire scenes or characters, and so on. In the Simulated Environment for Theatre, we reconceived the design to better support the affordances that are central to the task of the Director. For example, the character movements are now associated with a timeline rather than with the speeches, so that the character movement and the speeches are both attached to the proposed line of action on stage rather than to the text. We have also introduced the ability to judge the line of sight from any point in the audience by developing the system to load 3D scale models of actual stages and sets. Using the Unity game engine as the programming environment, we inherited the standard game controls for camera movement, so that quite sophisticated variations in perspective are possible. The user can, for instance, switch between cameras situated at different locations, making it possible to quickly see the stage from several angles. It is also possible to switch from the audience perspective to the perspective of any of the actors.

Our user study will be carried out early in 2010. We will ask six directors working in industry to try out the two interfaces and provide comments through a thinkaloud protocol. We will also obtain screen captures of these individuals working with the two different systems. Finally, considering pedagogical applications of these interfaces, we will work with a small group of graduate students in theatre education who will consider the affordances of the prototypes for teaching and learning about disciplinary theory and practice. We expect this process to provide additional insights into the choices of functionality for SET, as well as the features of Watching the Script that might be worth adapting for inclusion in the new system.

References

- Roberts-Smith, Jennifer, Gabriele, Sandra, Ruecker, Stan, Sinclair, Stéfan, Bouchard, Matt, DeSouza-Coelho, Shawn, Kong, Annemarie, Lam, David, Rodriguez, Omar (2009). 'The Text and the Line of Action: Re-conceiving Watching the Script'. *Proceedings of the INKE 2009: Research Foundations for Understanding Books and Reading in the Digital Age*. Victoria, BC, 23-24 October 2009.**

The Person Data Repository

Roeder, Torsten

roeder@bbaw.de

Berlin-Brandenburgische Akademie der
Wissenschaften

The *Berlin-Brandenburg Academy of Sciences and Humanities* (Berlin-Brandenburgische Akademie der Wissenschaften, BBAW) is the largest non-university research institution in the region. TELOTA (*The Electronic Life of the Academy*), an initiative for academically applied information technology, was launched here in 2002. The initiative supports the academy's projects by developing IT solutions for research work and digital publications.

The project "Construction of a repository for biographical data on historical persons of the 19th century" – short form: *Person Data Repository* – enhances the existing approaches to data integration and electronically supported research in biographies. It investigates connecting and presenting heterogeneous information on persons of the "long nineteenth century" (1789–1914). The project's aim is to provide a de-central software system for research institutions, universities, archives, and libraries that allows combined access on biographic information from different data pools.

The project is subdivided into three major fields: 1) conceptual design of an adequate data model, which embraces different methods and perspectives; 2) data exchange with national and international cooperation partners; and 3) development of a software solution based on an evaluated framework. The project is funded by the DFG (*Deutsche Forschungsgemeinschaft*, German Research Foundation). The work began in July 2009 with three academic staff members and three student assistants, and will continue for two years.

1. Data Modelling

To structure heterogeneous biographical data, the project pursues a novel approach, which was already presented in a talk at the workshop "Personendateien – Elektronisches Publizieren" in September 2009 in Leipzig (see link below). The approach does not define a person as single data record, but rather as compilation of all statements concerning that person. Thus, it is possible to display complementing as well as contradicting statements in parallel, which meets one of the basic challenges of biographic research.

In the above lecture by Niels-Oliver Walkowski, he notes: "Biographic research, understood as the creation of identifying narrations, performs semantic constructions, which were caused by a human, but which are not identical to it. The consequences are concurring narrations, polysemy and contingency, which are not an expression of lacking knowledge, but are due to the conditions of biographic research."

In order to satisfy different research approaches and perspectives, the smallest entity of the Person Data Repository is not a person, but a single statement on a person, which is named "aspect" in the data model. An aspect bundles references to persons, places, dates and sources. By proper queries it will be possible to create further narrations, whose first dimension is not necessarily a person, but possibly also a time span or a certain location. Those attained insights can enhance the knowledge on persons in turn.

Additionally, all aspects are connected to the corresponding source and to current identification systems respectively, like the LCCN or the German PND. Thus, scientific transparency and compatibility with existing and future systems is guaranteed.

2. Cooperation

As the Person Data Repository acquires data from partners, focusing on data organisation rather than conducting its own research, cooperation is an essential part of the project. The mission statement is guided by the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, which has also been signed by the *Berlin Brandenburg Academy of Sciences and Humanities*, and which promotes the exchange of scientific knowledge through digital media as well as transparency of sources and authors.

Initially, several projects of the BBAW were invited to share their data on the Person Data Repository. Amongst them are the *Marx-Engels-Gesamtausgabe* (MEGA, Complete works by Marx and Engels), the *Alexander-von-Humboldt-Forschungsstelle*, the *Protokolle des Preußischen Staatsministeriums* (Protocols of the Prussian Ministry of State), the *Berliner Klassik* (Classical Berlin) and the *Altmitgliederverzeichnis* (Index of Former Academy Members). Thus, an inventory of several thousand persons is already available; partnerships with further BBAW projects are planned in order to reach about 200,000 entries.

Parallel to this, partnerships with external institutions are being formed. These cooperations will range from sharing unstructured data to data exchange with existing repositories. Contacts with edition projects, image databases and repository

databases have already been formed and will be developed during the project's course. A basis for exchange and publication of the gathered person information should be delivered by Open Access oriented agreements, whereas individual arrangements can be made as well.

As it is intended to provide not only the Person Data Repository's contents, but also its infrastructure, it is also of interest for projects whose historical scope is focused outside the 19th century. In this case, it is possible set-up an infrastructure of the same type with freely configurable contents, as it were a sister repository.

3. Development

During the preliminaries for the technical realisation of the Person Data Repository, a list of established software packets has been created. An evaluation shows which of these packets will constitute the core component of the repository. As the software is provided also to other institutions, key elements of the evaluation are documentation, configurability, scalability, expandability and availability of interfaces.

A software which has already been utilized by BBAW projects for gathering person data is the "Archiv-Editor" (Archive Editor), which had been developed by the TELOTA initiative (see link below). This editor will play a central part as an editing tool in the Person Data Repository, and will be developed further according to the project's demands.

Along with the work on the repository and archive software, also the conversion of data is a part of the development field. Manual and automated methods will be utilized in the process. The atomization in single aspects is conducted in three steps: syntax analysis, index based structuring, and manual correction. The first data pool to convert was the person index of the Protocols of the Prussian Ministry of State, a completed project which provides an excellent starting basis, containing over 22,000 person entries. As the second major source for data on persons of the 19th century, the exhaustive indexes of the Complete Works of Marx and Engels have been chosen. Further projects at the academy, like the Alexander-von-Humboldt-Forschungsstelle and Classical Berlin, have placed their data pools at the disposal of Person Data Repository.

4. Perspectives

As the project aims at cooperation to a great extent, it is our wish to communicate with interested parties from all disciplines, in order to build up partnerships between our institutions. Also, we look forward to questions, remarks, proposals, and to exchanging

theoretical and technical approaches. A cooperative workshop for partners and interested parties is planned for autumn 2010.

Die *Berlin-Brandenburgische Akademie der Wissenschaften* (BBAW) ist die größte außeruniversitäre Forschungseinrichtung der Region. Im Jahr 2002 wurde hier TELOTA (*The Electronic Life Of The Academy*), eine Initiative für akademisch angewandte Informationstechnologie, ins Leben gerufen. Diese unterstützt die Akademievorhaben mit der Entwicklung informationstechnischer Lösungen für Forschungsarbeit und digitale Publikation.

Mit dem DFG-Projekt „Aufbau eines Repositoriums für biografische Daten historischer Personen des 19. Jahrhunderts“ – kurz: *Personendaten-Repositorium* – werden bisherige Ansätze der Datenvernetzung und elektronischen Biografik weiterentwickelt. Es erforscht anhand von Personeninformationen des „langen 19. Jahrhunderts“ (1789–1914), wie sich heterogene Datenbestände miteinander verbinden und präsentieren lassen. Ziel des Projektes ist die Bereitstellung eines dezentralen Softwaresystems, welches Lehr- und Forschungseinrichtungen, Archiven und Bibliotheken ermöglicht, biographische Informationen aus verschiedenen Beständen über einen gemeinsamen Zugang zu nutzen.

Das Projekt untergliedert sich in drei Teile: 1) Der Entwurf eines geeigneten Datenmodells, welches unterschiedlichen Perspektiven und Forschungsmethoden gerecht wird, 2) der Datenaustausch mit Kooperationspartnern im In- und Ausland, und 3) die Entwicklung einer Software-Lösung auf der Basis eines zu evaluierenden Framework. Das Projekt wurde von TELOTA bei der DFG beantragt und ist für die Laufzeit von zwei Jahren bewilligt worden. Im Juli 2009 wurde die Arbeit mit drei wissenschaftlichen Mitarbeitern und drei studentischen Hilfskräften aufgenommen.

1. Datenmodellierung

Zur Strukturierung heterogener biographischer Daten verfolgt das Projekt einen neuartigen Ansatz, der bereits in einem Vortrag auf dem Workshop „Personendateien – Elektronisches Publizieren“ im September 2009 in Leipzig vorgestellt wurde (siehe Link unten). Eine Person wird darin nicht als einzelner Datensatz definiert, sondern vielmehr als die Menge aller Aussagen, die zu ihr getroffen werden. Damit ist es möglich, sowohl sich ergänzende als auch sich widersprechende Aussagen

nebeneinander abzubilden, was grundlegenden Problemen biografischen Arbeitens Rechnung trägt.

In dem erwähnten Vortrag von Niels-Oliver Walkowski hieß es: „Biografisches Arbeiten verstanden als Erzeugung von identifizierenden Narrationen vollzieht semantische Konstruktionsleistungen, zu denen ein Mensch den Anlass gab, der aber nicht mit ihm zusammenfällt. Eine Folge sind konkurrierende Narrationen, Polysemie und Kontingenz, die nicht Ausdruck mangelnder Kenntnis, sondern den Voraussetzungen biografischen Arbeitens an sich geschuldet sind.“

Gerade also, um verschiedenen Forschungsansätzen und Perspektiven gerecht zu werden, ist die kleinste Dateneinheit des Personendaten-Repositoriums nicht eine Person, sondern eine einzelne Aussage zu einer Person, die in dem Datenmodell „Aspekt“ genannt wird. Ein Aspekt bündelt Bezüge zu Personen, Orten, Daten und einer Quelle. Dadurch wird es möglich sein, durch eine entsprechende Abfrage weitere Narrationen zu erzeugen, bei denen nicht unbedingt eine einzelne Person, sondern auch ein Zeitraum oder ein Ort die erste Dimension bilden könnte. Daraus gewonnene Erkenntnisse erweitern wiederum das Personenwissen.

Zudem werden die Aspekte einerseits mit den jeweiligen Quellen, andererseits mit geläufigen Identifikationssystemen, etwa mit PND und LCCN, verknüpft. Dadurch bleibt die wissenschaftliche Transparenz und die Kompatibilität mit bestehenden und zukünftigen Systemen gewährleistet.

2. Kooperationen

Da das Personendaten-Repositorium die Daten über seine Partner bezieht und sich selbst auf die Organisation der Daten konzentriert, anstatt eigene Datenbestände zu erarbeiten, ist Zusammenarbeit ein essenzieller Bestandteil des Projektes. Von richtungweisender Bedeutung sind dabei die Ziele der *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, die auch von der Berlin-Brandenburgischen Akademie der Wissenschaften unterzeichnet wurde und welche den wissenschaftlichen Austausch durch digitale Technik unter Gewährleistung von Quellen- und Autorentransparenz befördert.

Zunächst wurden einige Vorhaben der Berlin-Brandenburgischen Akademie der Wissenschaften dazu eingeladen, ihre Daten auf der zukünftigen Repositorien-Plattform verfügbar und verknüpfbar zu machen. Darunter fallen die *Marx-Engels-Gesamtausgabe*, die *Alexander-von-Humboldt-Forschungsstelle*, die *Protokolle des Preußischen Staatsministeriums* sowie das *Altmitgliederverzeichnis aus dem Akademiearchiv*.

Damit liegen bereits Daten zu mehreren zehntausend Personen vor; weitere Kooperationen mit Akademievorhaben sind in Vorbereitung, um ca. 200.000 Einträge zu erreichen.

Parallel dazu werden externe Institutionen als Partner herangezogen. Die Kooperationsmöglichkeiten reichen von der Übernahme unstrukturierter Datenbestände bis hin zum Austausch mit anderen Repositorien. So sind bereits Kontakte mit mehreren Editionsprojekten, Bilddatenbanken und Verbund-Datenbanken geknüpft worden, die im weiteren Verlauf des Projektes zu vertiefen sind. Die Grundlage für Austausch und Veröffentlichung der Personendaten schafft im besten Falle eine Vereinbarung im Sinne des *Open Access*, wobei auch individuell davon abweichende Verabredungen getroffen werden können.

Da beabsichtigt ist, nicht nur die Daten, sondern auch die Infrastruktur des Personendaten- Repositoriums zur Verfügung zu stellen, ist das Projekt auch für solche Vorhaben interessant, deren historischer Rahmen das 19. Jahrhundert nicht berührt. In diesem Fall kann eine Abmachung über die Einrichtung einer gleichartigen Infrastruktur mit selbst bestimmbaren Inhalten, also eine Art „Schwester-Repositorium“, geschlossen werden.

3. Entwicklung

Im Rahmen der Vorbereitungen für die praktische Umsetzung des Personendaten-Repositoriums wurde eine Liste etablierter Software-Pakete erstellt. Eine Evaluation zeigt, welches die Kernkomponente der Datenhaltung des PDR bildet. Da die Software auch anderen Projekten zur Verfügung steht, gehören Dokumentation, Konfigurierbarkeit, Skalierbarkeit, Erweiterbarkeit und Schnittstellen zu den wesentlichen Anforderungen der Evaluation.

Als bereits längerfristig genutzte Software zur Erfassung von Personendaten existiert innerhalb der BBAW der von der TELOTA-Initiative entwickelte Archiv-Editor (Link s. u.). Dieser wird im Rahmen des PDR eine zentrale Rolle als Werkzeug für die Eingabe von Personendaten spielen und wird entsprechend der veränderten Anforderungen weiterentwickelt.

Neben der Arbeit an der Repositorien- und Archivsoftware fällt auch die Konvertierung von Datenbeständen in den Entwicklungsbereich. Dabei werden sowohl manuelle als auch automatische Verfahren eingesetzt. Die Zerlegung der Biogramme in Einzelaspekte erfolgt zunächst anhand einer einfachen Syntaxanalyse, wird dann über Abkürzungs-, Orts- und Personenverzeichnisse tiefstrukturiert und zum

Abschluss manuell korrigiert. Begonnen wurde mit dem Personenregister der *Protokolle des Preußischen Staatsministeriums*, welche sich als abgeschlossenes Projekt und mit über 22.000 Kurzbiogrammen als hervorragende Ausgangsbasis anbot. Als zweite Quelle für Personendaten des 19. Jahrhunderts wurden die umfangreichen Register der *Marx-Engels-Gesamtausgabe* ausgewählt. Weitere Akademienvorhaben, etwa die *Alexander-von-Humboldt-Forschungsstelle*, haben ihre Daten ebenfalls bereits zur Verfügung gestellt.

4. Ausblick

Da unser Projekt in großem Maße auf Kooperationen setzt, ist es unser Wunsch, mit interessierten Projekten aus allen denkbaren Fachgebieten ins Gespräch zu kommen und Partnerschaften zu schließen. Ebenso freuen wir uns auf Fragen, Hinweise und Anregungen sowie auf den Austausch von theoretischen und technischen Lösungsansätzen. Ein Workshop, zu dem sowohl ähnliche Projekte, Kooperationspartner und interessiertes Fachpublikum eingeladen werden, ist für Herbst 2010 geplant.

Further information

- Website of the Person Data Repository <http://pdr.bbaw.de>
- The TELOTA Initiative of the BBAW <http://www.bbaw.de/telota>
- The "Archiv-Editor" (Introduction and Download) <http://www.bbaw.de/telota/projekte/personendatenbank-1/archiv-editor>
- "Personendateien" Workshop in Leipzig <http://www.saw-leipzig.de/aktuelles/personendateien>

Events

Workshop: Personen - Daten - Repositorien (Persons - Data - Repositories), 27th – 29th September 2010 Berlin-Brandenburg Academy of Sciences and Humanities: <http://pdr.bbaw.de/workshop>.

References

Bräse, J., Klump, J. (2007). 'Zitierfähige Datensätze: Primärdaten-Management durch DOIs'. *Rafael Ball, Wissenschaftskommunikation der Zukunft*. Jülich: Forschungszentrum Jülich. <http://books.google.com/books?id=kouJ09GQtbcC&lpg=PA159&ots=YuCzaRjSDM&dq=Zitiert%20hige%20Datens%20tze%3A%20Prim%20rdaten-Management%20durch%20DOI&lr=&pg=PA159#v=onepage&q=&f=false>.

Costa, Stefano (2010). *Open Data in Archaeology*. <http://blog.okfn.org/2010/02/25/open-data-in-archaeology/>.

Dallmeier-Tiessen, S., Dobratz, S., Gradmann, S., Horstmann, W., Kleiner, E., Pampel, H. (et al.) *Positionspapier Forschungsdaten*. <http://edoc.gfz-potsdam.de/gfz/13230>.

Dallmeier-Tiessen, Sunje, Pfeiffenberger, Hans (2009). 'Umgang mit Forschungsdaten in den Geowissenschaften - Ein Blick in die Praxis'. *Bibliothekartag 2009*. Erfurt: Berufsverband Information Bibliothek e.V.. http://www.opus-bayern.de/bib-info/volltexte/2009/699/pdf/dallmeier-tiessen_geowissenschaften.pdf.

DINI, Arbeitsgruppe "Elektronisches Publizieren" (2009). *Positionspapier Forschungsdaten*. Humboldt-Universität zu Berlin. <http://edoc.hu-berlin.de/series/dini-schriften/2009-10/PDF/10.pdf>.

Griese, B., Grieshop, H. R. (2007). *Biographische Fallarbeit*. Wiesbaden: VS Verlag für Sozialwissenschaften. <http://www.ulb.tu-darmstadt.de/tocs/177279664.pdf>.

Hackländer-von der Way, Bettina (2001). *Biographie und Identität*. Berlin. <http://dissertation.de>.

Henning, Tim (2009). *Person sein und Geschichten erzählen, Quellen und Studien zur Philosophie*. Berlin u.a.: de Gruyter.

Hermann, Elfriede (2003). *Lebenswege im Spannungsfeld lokaler und globaler Prozesse: Person, Selbst und Emotion in der ethnologischen Biografieforschung*. Münster: LIT.

Hoerning, Erika (2000). *Pierre Bourdieu: Die biographische Illusion*. Stuttgart: Lucius & Lucius.

Hoerning, Erika (2000). *Biographische Sozialisation*. Stuttgart: Lucius & Lucius.

Hutto, Daniel D. (2007). *Narrative and understanding persons*. Cambridge University Press. <http://books.google.de/books?id=peHYAAQAAQMAAJ>.

Hutto, Daniel (2007). *Framing Narratives*. Cambridge/New York: Cambridge University Press.

Key Perspectives Ltd.. *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability*. http://www.dcc.ac.uk/docs/publications/SCARP_SYNTHESIS.pdf.

- Kramer, Christine** (2001). *Lebensgeschichte, Authentizität und Zeit*. Frankfurt am Main u.a: Lang.
- Kripke, Saul** (1981). *Name und Notwendigkeit*. Frankfurt a.M.: Suhrkamp.
- Mackenzie, Catriona** (2008). *Practical identity and narrative agency*. New York: Routledge.

Marotzi, Winfried (2001). *Methodologie und Methoden der Biographieforschung*. Hohengehren: Schneider.

Moore-Gilbert, B. (2009). *Postcolonial life-writing: culture, politics and self-representation*. London/New York: Routledge.

NESTOR Arbeitsgruppe Grid/e-science und Langzeitarchivierung (2009). *nestor-bericht - Digitale Forschungsdaten bewahren und nutzen - für die Wissenschaft und die Zukunft*. Frankfurt am Main. <http://nbn-resolving.de/nbn:de:0008-2009071031>.

Neuroth, Heike, Jannidis, Fotis, Rapp, Andrea, Lohmeier, Felix (2009). 'Virtuelle Forschungsumgebungen für e-Humanities. Maßnahmen zur optimalen Unterstützung von Forschungsprozessen in den Geisteswissenschaften'. *Bibliothek, Forschung und Praxis*. **33** (2): 161-169. <http://www.reference-global.com/doi/abs/10.1515/bfup.2009.017>.

Schwiegelsohn, U. (2009). 'Grids als neue Komponenten des Integrierten Informationsmanagements'. *Praxis der Informationsverarbeitung und Kommunikation*. **32** (1): 29-32. <http://www.reference-global.com/doi/pdfplus/10.1515/piko.2009.006>.

Wettlaufer, Jörg. 'Personendateien. Workshop der Arbeitsgruppe Elektronisches Publizieren der Union der deutschen Akademien der Wissenschaft - H-Soz-u-Kult / Tagungsberichte'. *H-Soz-u-Kult*. <http://hs-ozkult.geschichte.hu-berlin.de/tagungsberichte/id=2806&count=126&recno=9&sort=datum&order=down&search=presse&epoch=22>.

Structured and Unstructured: Extracting Information from Classics Scholarly Texts

Romanello, Matteo

matteo.romanello@kcl.ac.uk
King's College London

The poster presents an ongoing PhD research project that applies Digital Humanities to Classics. The project is focussed on the extraction of information from modern scholarly texts (i.e. "secondary sources"), namely all the modern publications about ancient works written in Greek or Latin (being our so-called "primary sources"). The project addresses both the problem of extracting information from scholarly texts in an automatic and scalable way, and that of providing users (i.e. scholars) with advanced and meaningful entry points to information rather than just search engine-like functionalities over an electronic corpus of texts.

1. Background

A currently ongoing project such as the Million Book Library drew considerable attention to the characteristic features of the next generation of digital libraries and to the consequences of a change of scale on the practice and the results of the research itself (Crane 2006). This is a big chance for Humanities and Classics particularly given the extended "shelf-life" of humanities relative to scientific publication. However, once those secondary sources are digitized this does not mean that information contained within them will be immediately accessible. Issues that we need to address concern the accuracy of our electronic resources, such as encoding of Greek text, inaccurate OCR transcription due to low image quality, problem of missing pages (Boschetti 2009), as well as the scalability of ways to provide access to information, given that we cannot afford to correct manually every scanned page.

2. Motivations

Secondary sources are without any doubt intrinsically valuable for Classicists, as they shape the scholarly discourse of the discipline. Printed citations and references – and even mentions of names or geographical places – can be considered as being already a form of hypertext as they virtually create links between texts. However in the currently

available digital libraries – except for the Perseus Digital Library¹ – primary and secondary sources are scarcely interconnected, despite the fact that one of the main advantages of digital libraries is the way in which they represent the hypertextual nature of text collections.

In particular, Classics scholars are interested in named entity mentions inside texts, as is reflected by the widespread use of different kinds of indexes and concordances in this field that basically organize in a systematic manner references to text passages when a given entity is mentioned. Translating the problem into computational terms, we are faced with the task of automatic entity extraction from a corpus of unstructured texts to develop a discipline-specific system of semantic information retrieval.

3. Related Work

In addition to all the unstructured information being made available on the web, several projects in the Humanities have produced over the last decade an increasing amount of structured information that was then stored in a wide range of data formats (i.e. databases, XML files, etc.). The approach we undertake is to reuse information contained within those structured data sources as training data for a supervised system that extracts semantic information from an unstructured corpus of texts. A likely approach is suggested for instance by a research recently conducted by IBM to investigate the automatic creation of links between structured data sources (i.e. database containing product information) and unstructured texts (i.e. emails of complaint about products) (Bhide et al. 2008). More generally, the problem implied by our approach of determining when two bits of information refer to the same entity has been thoroughly explored in the AI (Artificial Intelligence) field (Li et al. 2005).

4. Method

The very first phase of the project is devoted to the task of building our corpus of unstructured texts. So far we considered two corpora of texts: the papers contained in the open archive Princeton/Stanford Working Papers in Classics² (Josiah Ober et al. 2007) and the articles published by the journal Lexis³ available online under an open access policy. Although the texts are already available in electronic format, some pre-processing is needed – particularly for the sequences of Greek text contained – before we can start extracting information.

In the second phase we integrate different structured data sources into a single knowledge base. As far as this task is concerned, it is possible to observe at least two main categories of lack of

interoperability. The first category consists of cases where entities that are similar from an ontological perspective (e.g. a geographical place name) are encoded using different data structures (i.e. the same place name could be encoded using elements belonging to different XML dialects). The second category covers cases where chunks of information common to more than one collection are described with different degrees of depth and precision. For instance, the name “Alexandria” inside an inscription is just marked up as a place name where instead a collection of geographical data offers many details for the city of Alexandria such as coordinates, orthographical variants of the name, denomination in different languages etc. For this purpose we mean to apply high level ontologies to aggregate information related to the same entity but spread over different data sources. Among the most suitable ontology vocabularies are worth to be mentioned FOAF,⁴ CIDOC CRM,⁵ FRBRoo⁶ and YAGO.⁷

At a further stage we are taking into account how to automatically extract information from the corpus. We are mainly interested in extracting: 1) named entities; 2) bibliographic references; 3) canonical references, namely references to ancient texts expressed in a concise form and characterized by a logical reference scheme (e.g. based on references to books or lines of a work instead of page numbers).

The very first step in named entities processing is the recognition and identification of named entities within texts. Once identified the named entities should be classified and then disambiguated on the basis of the context. This task will be accomplished through the comparison of semantic spaces using methods and algorithms developed in the field of Latent Semantic Analysis (LSA) (Rubenstein & Goodenough 1965; Sahlgren 2006). In particular the semantic spaces of all the contexts where a given named entity appear will be compared with each other in order to determine which resources are really referring to the same entity.

For the information extraction task we use mainly tools based on machine learning methods, such as Conditional Random Fields (CRF) or Support Vector Machine (SVM). Some of them are already available as open source software and just need to be trained to work with our data, while others are being specifically developed such as a Canonical Reference Extractor (Romanello et al. 2009). Information contained in the knowledge base is meant to be used as training material for those software components.

5. Further Work

This ongoing project is expected to prove a scalable approach to extract entity references from

unstructured corpora of texts, such as OCRed materials. In addition to this, it will prove how to reuse several data sources of structured information to train text mining components and it will show to what extent those data sources can be made interoperable with each other. Finally we plan to evaluate with some experts how effective such an information retrieval system is in helping Classicists with their research.

Acknowledgement

This research project is supported by an AHRC (Arts and Humanities Research Council) award.

References

3. <http://lexisonline.eu/>
 4. <http://www.foaf-project.org/>
 5. <http://cidoc.ics.forth.gr/>
 6. http://cidoc.ics.forth.gr/frbr_inro.html
 7. <http://www.mpi-inf.mpg.de/yago-naga/yago/>
- Bhide, M. et al.** (2008). 'Enhanced Business Intelligence using EROCS'. *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on Data Engineering, 2008.* Pp. 1616-1619. <http://ieeexplore.ieee.org/iel5/4492792/4497384/04497635.pdf?tp=&arnumber=4497635&isnumber=4497384>.
- Boschetti, F.** (2009) (2009). 'Improving OCR Accuracy for Classical Critical Editions'. *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*. Springer.
- Crane, G.** (2006). 'What Do You Do with a Million Books?'. *D-Lib Magazine*. **12(3)**. <http://www.dlib.org/dlib/march06/crane/03crane.html> (accessed March 19, 2009).
- Josiah Ober et al.** (2007). 'Toward Open Access in Ancient Studies: The Princeton-Stanford Working Papers in Classics'. <http://www.atypon-link.com/ASCS/doi/abs/10.2972/hesp.76.1.229> (accessed July 15, 2009).
- Li, X., Morie, P., Roth, D.** (2005). 'Semantic integration in text: from ambiguous names to identifiable entities'. *AI Mag.* **26(1)**: 45-58. <http://portal.acm.org/citation.cfm?id=1090494> (accessed March 12, 2010).
- Romanello, M., Boschetti, F., Crane, G.** (2009). 'Citations in the digital library of classics: extracting canonical references by using conditional random fields'. *NLPiR4DL '09: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Morristown, NJ, USAAssociation for Computational Linguistics, pp. 80–87.

Notes

1. <http://www.perseus.tufts.edu/hopper/>
2. <http://www.princeton.edu/~pswpc/>

Original, Translation, Inflation. Are All Translations Longer than Their Originals?

Rybicki, Jan

jkrybicki@gmail.com

Pedagogical University, Krakow, Poland

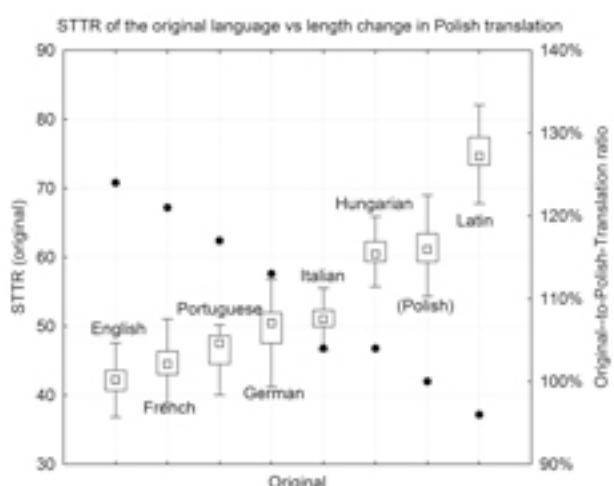
It is a truth almost universally acknowledged, at least among translator service providers, that some languages take fewer words to express the same thing than some other languages, to the extent that translator remuneration is often calculated accordingly. To further paraphrase Jane Austen and John Burrows: this truth is so well fixed in the minds of the general translating community that the scant reports pointing to the contrary – or, at least, to a possibility that this effect might be exactly contrary to expectations – are either ignored or appear in the wrong journals (Rybicki, 2006). While this problem is not entirely ignored by traditional translation studies, it is usually dealt with as an aside in publications where this discipline meets corpus linguistics to define and study translator style (Baker, 1993, 1996, 2000), or applied to no more than two languages, very few texts and, more often than not, small sample sizes (Englund Dimitrova, 1994, 2003, Pápai, 2004), or oriented to point out differences between two translations rather than original and translation (Rybicki, 2009). Theoretical considerations are just as unsatisfactory. Differences in the level of inflection of the two languages are usually seen as the reason for the differences in length between the native and the foreign version of the same text; the rare exception, i.e. a more or less positive statement on the subject, has been made by George Steiner: “translations are inflationary” (Steiner, 1978) in a discussion of explication, one of the so-called translation universals (Baker, 1993, 1996). Still, while explication is a mechanism that certainly does involve using more text in the target language to denote less text in the original, it is not clear whether Steiner had any specific textual unit in mind that would undergo inflation – as opposed to another possibility, the inflation of meaning.

Indeed, it is even less clear whether mere difference in the number of words – the first and reflexive approach most stylometrists would take – between a novel in one language and another novel, the former's translation, is at all of any scholarly interest; it is quite possible that what matters more is the increase (or

decrease) in the number and/or the length of, say, sentences. Even then, however, the differences could be a simple consequence of the divergent linguistic systems and the whole problem should be left at that.

It is almost a tradition that, faced with such theoretical quandaries, members of our community turn to empirical practicalities, to experiment – and this is exactly what this paper does. Using a series of fairly extensive bilingual corpora or, simply speaking, combinations of original and translation (and, in some cases, another, and yet another translation of the same text) in a variety of source and target languages, the study compares the sizes, establishes their patterns and their statistical significance (with z-scores). The corpora in question include: English translations of Polish novels by Henryk Sienkiewicz; Polish translations of American, English, French, German and Italian prose (including the interesting sub-corpora of translations of Tolkien and of translations by the author of this paper); French and Polish translations of Shakespeare; Polish and English translation of Latin prose, Portuguese translations of English prose.

The results do not paint a uniform picture. While expected general trends can be observed in size variation between pairs of languages, the discrepancies in “inflation rate” between certain rival translations into the same language at times hide any stable “language-to-language” effect. This effect has been hypothetically ascribed at first to differences between inflected (agglutinative) and analytic languages. While this would be difficult to prove, at the same time – barring such extreme cases of translator logorhea as W.S. Kuniczak's famously overflowing translation of Henryk Sienkiewicz's historical romances, where the translation-to-original ratio reaches the vertiginous heights of 170%, the record value in the entire project – some correlation has been observed not so much between the general degree of inflection of a given language as between standardized type-token ratios in each of the studied individual-language corpora. Thus, although it would be too much to say that STTR is a good measure of a language's inflection, the general trend in STTR ranges observed in each of the corpora used in this study corresponds fairly well to the *inverted* order of languages exhibiting difference between original and translation (see Figure below): translations into English tend to be longer than their Polish originals; Polish translations are shorter than original English novels; most translations of Latin prose tend to be longer than the originals, and so forth. With an important caveat: it only takes an overambitious, overzealous or pathologically lazy translator, or an unscrupulous publisher, to alter this pleasant image beyond recognition.



Standardized Type-Token Ratio (Box & Whisker) and Original-to-Translation Ratio (Scatterplot) in Selected Prose Corpora

References

- Baker, M.** (1993). 'Corpus linguistics and translation studies: Implications and applications'. *Text and Technology: In honour of John Sinclair*. Baker, M., Francis, G., Tognini-Bonelli, E. (eds.). Amsterdam: John Benjamins, pp. 17-45.
- Baker, M.** (1996). 'Corpus-based translation studies: The challenges that lie ahead'. *Terminology, LSP and Translation: Studies in language engineering, in honour of Juan C. Sager*. Somers, H. (ed.). Amsterdam: John Benjamins, pp. 175-186.
- Baker, M.** (2000). 'Towards a methodology for investigating the style of a literary translator'. *Target*. **12**: 241-266.
- Pápai, V.** (2004). 'Explicitation – A universal of translated text?'. *Translation Universals. Do they exist?*. Mauranen, Kujamäki (eds.). Amsterdam – Philadelphia: John Benjamins, pp. 143-164.
- Englund Dimitrova B.** (1994). *Statistical Analysis of Translations (On the basis of translations from and to Bulgarian, Russian and Swedish)*. Scandinavian Working Papers on Bilingualism. V. 9, pp. 87-103.
- Englund Dimitrova B.** (2003). 'Explicitation in Russian-Swedish translation: sociolinguistic and pragmatic aspects'. *Swedish Contributions to the Thirteenth International Congress of Slavists, Ljubljana, 15-21 August 2003*. Englund Dimitrova B., Pereswetoff-Morath, A. (eds.). Lund: Lund University, pp. 21-31.
- Rybicki, J.** (2006). 'Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations'. *Literary and Linguistic Computing*. **21(1)**: 91-103.
- Rybicki, J.** (2009). 'Liczenie krasnoludków. Trochę inaczej o polskich przekładach trylogii Tolkiena'. *Po co ludziom krasnoludki?*. Warszawa, 2009.
- Steiner, G.** (1978). *After Babel. Aspects of Language and Translation*. Oxford: Oxford University Press V. 253, reprinted, 1992.

A Platform for Cultural Information Visualization Using Schematic Expressions of Cube

Saito, Shinya

saitos@fc.ritsumei.ac.jp

Postdoctoral Fellow, Digital Humanities Center for Japanese Arts and Cultures Ritsumeikan University, Kyoto, Japan

Ohno, Shin

shinohno@gmail.com

Research Assistant, Digital Humanities Center for Japanese Arts and Cultures Ritsumeikan University, Kyoto, Japan

Inaba, Mitsuyuki

inabam@sps.ritsumei.ac.jp

Digital Humanities Center for Japanese Arts and Cultures Ritsumeikan University, Kyoto, Japan

In recent years, people have tended to be overwhelmed by a vast amount of information in various contexts. Therefore, arguments about "Information Visualization" as a method to make information easy to comprehend are more than understandable.

This paper will argue the method of visualization of vast amount of information using 3-D viewer, and we will introduce an environment called KACHINA CUBE (KC) that can visualize various information using a "cube". We have introduced KC in DH2009 (Saito, Inaba and Ohno, 2009). Then, we limited the use of KC to geographic information. But now, KC can adopt not only geographic information but also various events. In this paper, we introduce KC's new design and function.

1. Design Concept of KC

1.1. Visualization Design

The most important thing in this research is to develop a Web system to integrate a large quantity of fragmentary information and to construct a method for visualizing a "scheme of things". For this purpose, we need to place all fragmentary information in the same context. Moreover, we have to come up with a way to put various information in a cube.

We decided to design KC in three dimensions, two dimensions for geographical information and

another one for temporal information (see Fig. 1). In this virtual 3D space (CUBE model), users can post formal and informal story fragments and can spin the cube (see Fig. 2). Among them, we call formal ones history fragments, and informal fragments story ones. KC also supports researchers to make linkages among fragments in periodical or logical order. We call a set of cultural fragments a storyline.

We are developing KC's search-engine for the most appropriate word. A fragment highlighted means that it includes the search term. Moreover, if there is more than one appropriate fragment, a line is drawn among those fragments. This function helps to find a hidden context or story among fragmented information. When words are searched and some words have a logical conjunction, multiple lines are drawn, as well as color-coded, based on the number of matched words. This search function thus makes visible the degree of connection among fragments.

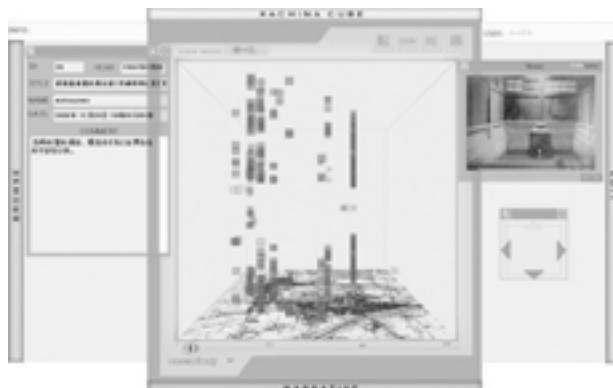


Fig 1. Application of CUBE Model for geographic information

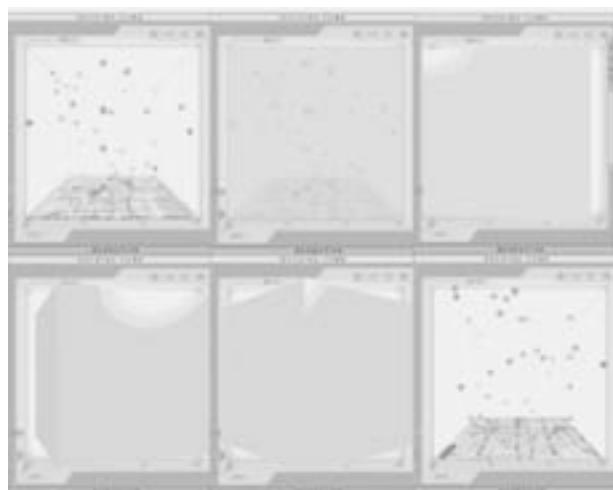


Fig 2. Rotation of cube

1.2. Sharing Design

W3C puts tremendous efforts to create standardized frameworks for Web, and researchers in digital humanities regard semantic web technology as one of the key research fields. This kind of technology gives us various chances to share data for other use. We

believe archived cultural data should be standardized to fit in this framework, which allows users to access data and utilize them in various platforms.

We apply RDF/OWL to define our data. Its extensive and flexible definition is suitable for our system and motivates other researchers to access our data (Bray, 2001).

1. History fragment class: Objective information in textbook or dictionary
2. Story fragment class: Subjective information such as oral history
3. Storyline class: Aggregate of historical and story fragments based on a specific context
4. Geography class: Geographical information of the historical and story fragments
5. Temporal class: Time when the incidents told in historical and story fragments occurred

1.3. Conceptual map Design

KC can adopt not only geographic information but also various events. In order to handle non-geographical information, KC supports conceptual maps. Moreover, the way in which a conceptual map is made should be considered. A conceptual map is a figure which expresses a development of story or event. We decided to apply the Trajectory Equifinality Model (TEM) to the process of making conceptual map. TEM is a theory to capture a certain phenomenon from view point of "time" and "process" (Valsinar and Sato 2006).

Moreover, KC is a web-based application that is built on the client-server architecture. In this system, client side application is implemented using ActionScript. On the other hand, server side application is developed by using PHP (PHP:

Hypertext Preprocessor). KC adopts MySQL as a relational database management system.

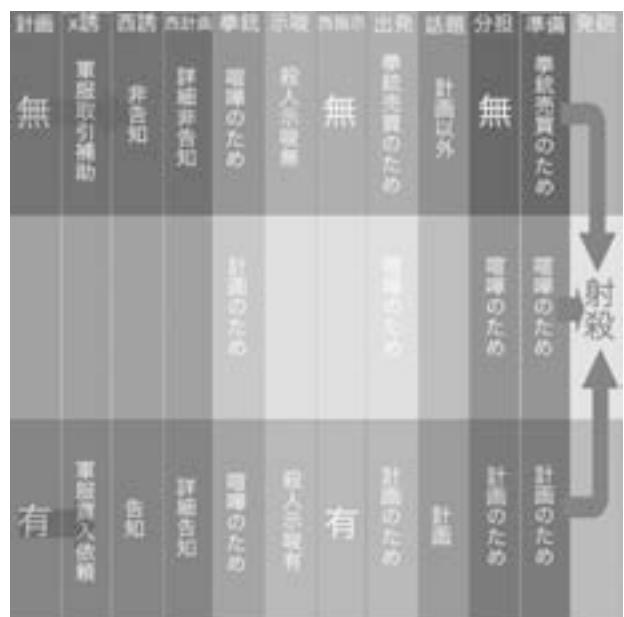


Fig 3. Conceptual map of a murder case

Below is an example of making a conceptual map using TEM. Fig.3 is a conceptual map to visualize a process of investigation and trial of a certain actual murder case. This conceptual map expresses both the process of the murder case itself and its trial. The map's 12 columns represent development of this murder case; and 3 rows, the statements in the court of the defendant and prosecution. The defendant's statements are set on the top, while that of prosecution on the bottom. Moreover, statements that are pertinent to neither the top nor the bottom are set in the middle.

Fig.4 shows a result of this visualization with KC. This system is suitable to grasp the perspective on complicated cases which include multiple statements, stories, or contexts.

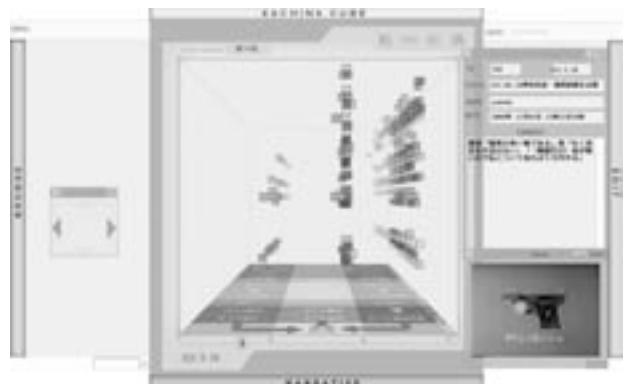


Fig 4. Visualization of the possess of trail

All that has to be done is to prepare conceptual maps, after which this method can be applicable to numerous cases in the humanities fields. For instance, by using this method, we are trying

to visualize argumentation about certain literary efforts.

Moreover, using visualization of search results function, we tried to find fragments which include "拳銃(gun)", "お金(money)", "警察(police)". As a result, three lines colored here in red, blue, and green appear (see Fig.5).

The red line means fragments which include "gun"; the blue, fragments with both "gun" and "money"; and the green, fragments which include all of "gun", "money" and "police". That is, the green line which includes all of the three shows the strongest tie.

Using this function allows not only us to extract fragments based on the given words, but also to detect some connections among information as well as their strength. Because of this function, it is possible to conduct text-mining through visual interface.

1.5. Conclusions and prospects

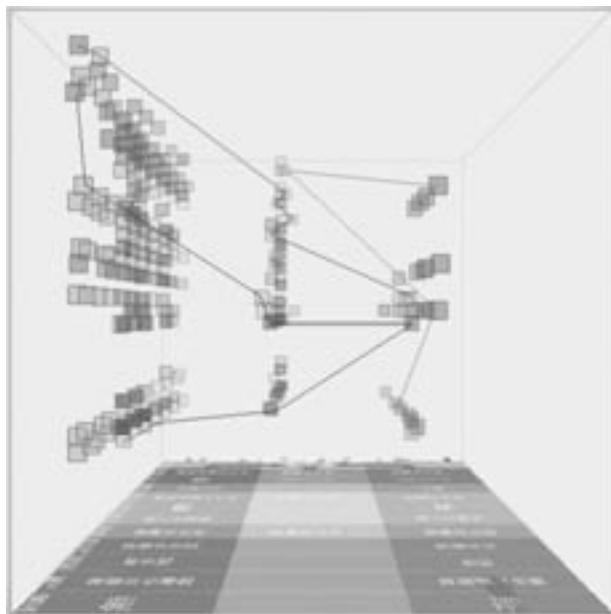


Fig 5. Visualization of search results

As a result, our system demonstrates a lot of potential for research in various fields, which we have to prove by developing further this software with applications, as well as examining it in more humanities case studies. Our KC is significantly different from these previous Web systems in the following three points:

1. Adoption of CUBE model (a 3D-viewer that combines the map with the timeline);
2. Implementation of a user interface suitable to contain a vast amount of information;
3. Implementation of analysis components for narratives and oral histories

We are trying to implement some new functions now, among which the most drastic is the "nesting cube" structure. All fragments in the KACHINA CUBE are defined as independent cubes in this structure. Furthermore, we can make a recursively-defined cube severalfold by using this function. Because of this, a large amount of information can be organized hierarchically.

For example, "England" could be set as a cube on the highest layer (root layer). Moreover, a cube of "Liverpool" can be set in "England". In "Liverpool", a cube of "The Beatles" might be attached. Therefore, it may include a cube of "John Lennon". Each cube can include any piece of information of the same level as a cube. This architectonics can change ways of organizing and browsing information fundamentally and support visualizing of a "scheme of things" very well. At the same time, we hope this architectonics will contribute greatly to Digital Humanities.

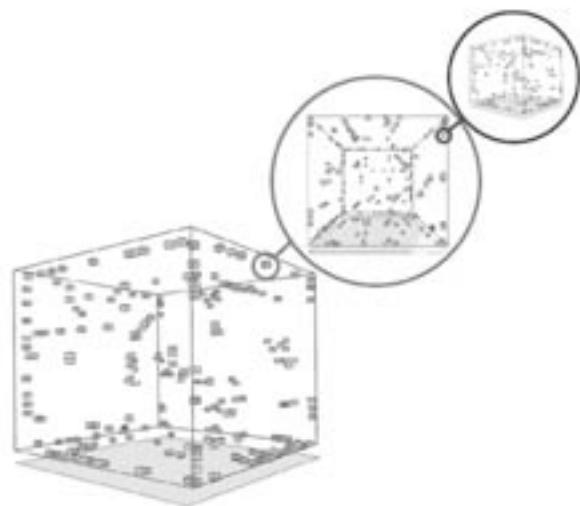


Fig 6. Nesting cube structure

References

- Saito, S., Ohno, S., Inaba, M.** (2009). 'Collective Culture and Visualization of Spatiotemporal Information'. *Proceedings of Digital Humanities 2009*. University of Maryland, USA, 22-25 June 2009, pp. 248-250.
- Bray, T.** (2001). *What is RDF?*. <http://www.w3.org/pub/a/2001/01/24/rdf.html> (accessed 14 November, 2008).
- Valsiner, J., Sato, T.** (2006). 'Historically Structured Sampling (HSS): How can psychology's methodology become tuned in to the reality of the historical nature of cultural psychology?'. *Pursuit of meaning. Advances in cultural and cross-cultural*

psychology. Straub, J., et al. (eds.). Bielefeld:
Transcript Verlag, pp. 215-251.

Generation of Emotional Dance Motion for Virtual Dance Collaboration System

Seiya, Tsuruta

seiyaimg.is.ritsumei.ac.jp

Graduate School of Science and Engineering,
Ritsumeikan University, Japan

Woong, Choi

Global Innovation Research Organization,
Ritsumeikan University, Shiga, Japan

Kozaburo, Hachimura

Department of Media Technology, College of
Information Science and Engineering, Ritsumeikan
University, Shiga, Japan

Measurement of body motion using motion capture systems has become widespread in the fields of entertainment, medical care, and biomechanics research.

In our laboratory, we are undertaking research on the application of digital archiving and information technology to dancing [Hachimura, 2006]. For example, quantitative analysis of traditional dance motion [Yoshimura et al., 2006] and 3D character animations of traditional performing arts using virtual reality [Furukawa et al., 2006]. Recently, we have measured many kinds of dance motion, not only Japanese traditional dances, but also contemporary street dances.

Dance collaboration system is one of the typical collaboration systems based on body motion. In our laboratory, we are developing a Virtual Dance Collaboration System [Tsuruta et al., 2007]. Live dancer's motion is captured by optical motion capture system, and the dance collaboration system recognizes it in real-time. A virtual dancer responds to the live dancer's motion. The live dancer performs a dance to the music, and a virtual dancer reacts with a dance by using the motion data stored in a motion database. It is desirable to generate a virtual dancer's motion according to the live dancer's emotion or music emotion.

In this paper, we describe a method to generate emotional dance motions by modifying a standard dance motion which is stored in a database.

1. Overview of the Virtual Dance Collaboration System

A configuration of the system is shown in Figure 1. Our proposed system provides users with collaboration with virtual dancers through dancing. The collaboration system consists of an optical real-time motion capture and an immersive virtual environment system. An optical motion capture system is able to measure body motion precisely and in real-time. Immersive virtual environment provides users stereoscopic display and feeling of immersion.

The system has three sections: "Motion processing section", "Music processing section" and "Graphics processing section".

In the "Motion processing section", the system recognizes a live dancer's motion in real-time, and determines a virtual dancer's reactive motion. In the "Music processing section", the system extracts emotional information from music in real time. In the "Graphics processing section", the system regenerates a virtual dancer's motion by using extracted emotional information from music, and the system displays 3DCG character animation by using immersive virtual environment.

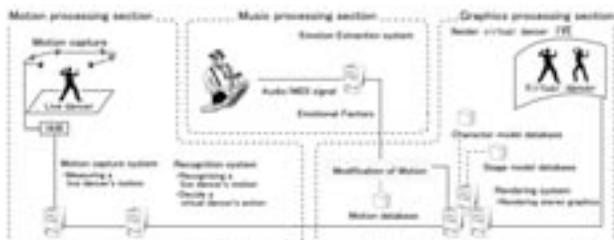


Fig. 1. Configuration of Virtual Dance Collaboration System

2. Generation of Emotional Dance Motion

For our system, it is necessary to generate a virtual dancer's emotional motion in real-time. We developed a system named Emotional Motion Editor (EME).

Feature \ Emotion	Speed of Motion	Height of Waist	Vertical Motion Range of Waist	Body Space
Neutral (Standard)	STD	STD	STD	STD
Passionate	Fast (+ +)	Little Low (-)	Wide (+ +)	Large (+ +)
Cheerful	A little fast (+)	Little High (+)	Same as STD	A little large (+)
Calm	A little slow (-)	Same as STD	Little small (-)	Small (- -)
Dark	Slow (- - -)	Low (- - -)	A little wide (+)	A little large (+)

Table 1 - Motion features appeared on each human emotion

Feature \ Emotion	Speed of Motion	Height of Waist		Vertical Motion Range of Waist		Body Space		
		joint	B	joint	n	joint	a	B
Passionate	1.1	knee	-35.0	knee	3.0	waist	2.0	5.0
Cheerful	1.1	knee	10.0	knee	1.0	waist	1.5	3.0
Calm	0.8	---	---	knee	-1.0	waist	1.0	<3.0
Dark	0.7	knee	-30.0	knee	1.5	waist	1.5	3.0
						elbow	2.0	---

Table II - Parameters used for modifying motion data

(α : coefficient β :bias)

2.1. Emotional Motion Editor

The EME generates emotional dance motions by modifying the original motion data by changing the speed of motion or altering the joint angles interactively. To generate an emotional motion, a function of changing the size of motion is implemented within the EME.

To generate virtual dancer's emotional motions in real-time, we need a simple method which can calculate with few order. For this purpose, we employ a method altering the interior angle of two connecting body segments as shown in Figure 2.

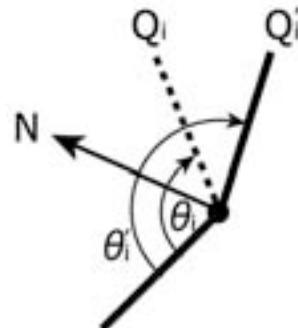


Fig. 2. Altering the joint angle

Q_i and Q'_i is an original vector and a vector after rotating respectively. Where θ_i is an original angle, and θ'_i is an after rotating angle respectively. The rotation matrix is calculated by using equation (1).

$$Q'_i = R_N(\theta'_i - \theta_i)Q_i \quad (1)$$

$$N = P_i \times Q_i \quad (2)$$

Where $R_N(\theta'_i - \theta_i)$ matrix for rotation about vector N . N is a normal vector represented by equation (2). Where \times means outer product.

Change the size of the motion is indicated by equation (3).

$$\theta'_i = \bar{\theta}_i + \alpha(\theta_i - \bar{\theta}_i) + \beta \quad (3)$$

$$\bar{\theta}_i = \frac{\sum_{t=-k}^k \theta_{i+t}}{2k+1} \quad (4)$$

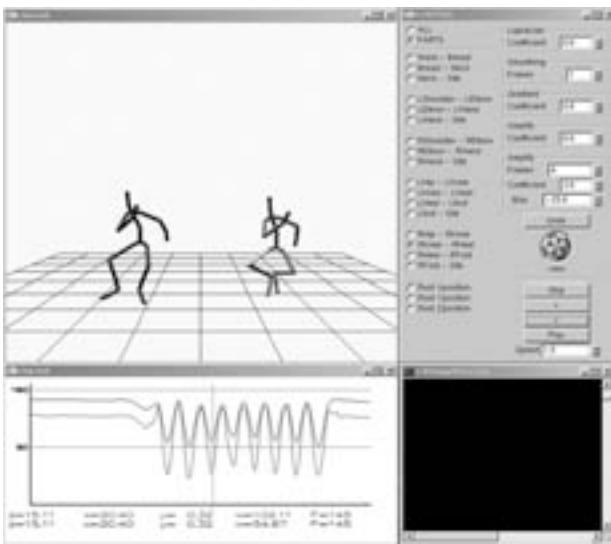


Fig. 3. Screen shot of Emotional Motion Editor

Constant α and β are coefficient for amplification and bias respectively. Where θ'_i is amplified angle, $\bar{\theta}_i$ is average of angles in sliding window at frame i as shown in equation (4). k is a half of the size of the sliding window.

A screen shot of the EME is displayed in Figure 3. A character model on the left shows an original dance motion. A model on the right shows generated emotional motion after modification.

2.2. Relation between Emotion and Body Motion

We examine the correlation between emotion and body motion in dancing by interviewing the dancer. We employ 5 kinds of emotions (Neutral, Passionate, Cheerful, Calm, Dark). "Neutral" is a standard motion. Motion features appeared on each human emotional motion are shown in Table I. We then obtained parameters empirically with a dancer. Parameters used for generating emotional motions are shown in Table II. Figure 4(a) shows an example of motion modification. A thin line in Figure 4(b) shows an original graph of angle variation of the right knee. The thick line shows a modified graph. In this case, α and β was given 3.0 and -25.0 respectively.

3. Experiments

We generate 4 kinds of emotional motions by using EME according to Table II.

To evaluate generated emotional motions, we conducted 2 types of assessment experiment by using questionnaire survey.

3.1. Method of Experiment

Experiment 1

Experiment 1 is a comparison between neutral standard motion and artificially generated emotional motions.

Experiment 2

Experiment 2 is a comparison between motion-captured emotional motions and artificially generated emotional motions.

For the experiment, we used 9 kinds of motions as the following:

- 4 Emotional motions (performed by dancer)
- 1 Standard motion (performed by dancer)
- 4 Artificial emotional motions (generated by EME)

3.2. Result of Experiments

The results of Experiment 1 are shown in Figure 5. This figure shows score averages, standard deviations and significant differences by the t-test. Black circles show standard motions, triangles show generated artificial emotional motions. As shown in Figure 5, all kinds of scores except "Calm" are higher than standard motion. We found that the respondents receive an impression of each emotion through artificial emotional motion.

Figure 6 indicates the results of Experiment 2. White circles show motion-captured emotional motions, and triangles show generated artificial emotional motions by using EME. As a result of the t-test, there is no significant difference. We found that respondents received similar impressions of emotional motions from generated artificial emotional motions. We verified that our EME system is effective in generating emotional motions.

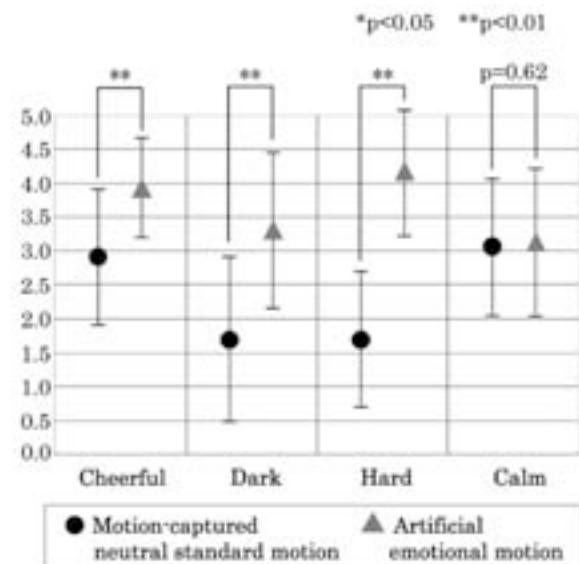


Fig. 5. Experimental result 1

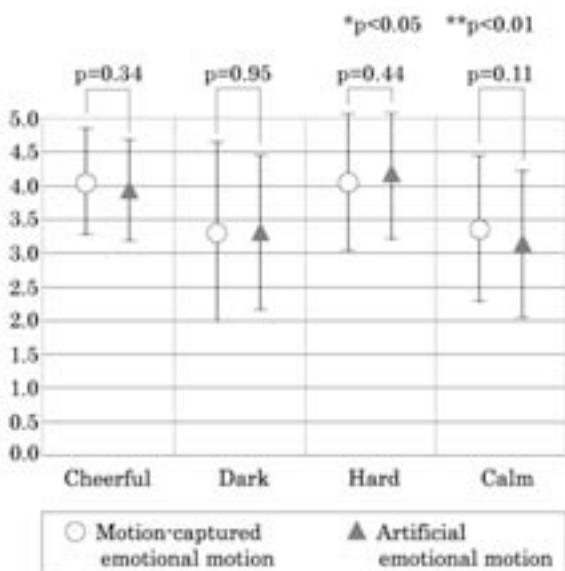


Fig. 6. Experiment result 2

4. Conclusion and future works

In this paper, we described a method to generate emotional dance motions by modifying the standard dance motion.

To generate emotional motions, we developed the Emotional Motion Editor. We conducted two experiments to evaluate generated emotional motions. As a result, we confirmed that EME can generate emotional motions by altering motion speed and joint angles.

As future work, implementation of the motion processing section and the motion modification function is necessary in the Virtual Dance Collaboration System.

Acknowledgements

This research has been partially supported by the Global COE Program “Digital Humanities Center for Japanese Arts and Cultures”, the Grant-in-Aid for Scientific Research No.(B)16300035, all from the Ministry of Education, Science, Sports and Culture. We would like to give heartfelt thanks to Prof. Y. Endo, Ritsumeikan Univ. whose comments and suggestions were of inestimable value for our research. We would also like to thank Ms. Gotan and Mr. Morioka who support many experiments.

References

- Furukawa, K. et al.** (2006). CG Restoration of Historical Noh Stage and its use for Edutainment. *Proc. VSMM06*. Pp. 358-367.
- Hachimura, K.** (2006). 'Digital Archiving of Dancing'. *Review of the National Center for Digitization (Online Journal)*.

51-66. <http://www.ncd.matf.bg.ac.yu/casopis/08/english.html> (accessed 12 March 2010).

Tsuruta, S. et al. (2007). 'Real-Time Recognition of Body Motion for Virtual Dance Collaboration System'. *Proceedings of 17th International Conference on Artificial Reality and Telexistence (ICAT 2007)*. 2007, pp. 23-30.

Yoshimura, M. et al. (2006). 'Analysis of Japanese Dance Movements Using Motion Capture System'. *Systems and Computers in Japan*. **No.1**: 71-82, Translated from Densi Joho Tsushin Gakkai Ronbunshi, Vol. J87-D-II, No.3.

“You don't have to be famous for your life to be history”: The Dusenberry Journal and img2xml

Smith, Natasha

nsmith@email.unc.edu

University of North Carolina at Chapel Hill, USA

Cayless, Hugh

hugh.cayless@nyu.edu

New York University, USA

This poster presentation will describe a project currently underway under the auspices of the Documenting the American South digital publishing program. It has been funded by grants from the US National Endowment for the Humanities (Digital Humanities Start-Up Grants program) and UNC Chapel Hill. The project is centered around the journal of a 19th century student at UNC named James Dusenberry and aims to use innovative web-based technology to present the journal and to create modules of supplementary material around it to provide insight into Dusenberry's world.

1. Background

The journal that forms the basis of our project was written by James Lawrence Dusenberry (1821-86) during the 1841-42 academic year. Dusenberry, the son of Lydia Davis (1797-1857) and planter Henry Rounsville Dusenberry (1794-1852) of Lexington, North Carolina, entered the University of North Carolina (UNC) in 1839. Sometime before graduating, he began copying out poems and lyrics to popular songs that he admired, and in July 1841 he began “Records of My Senior Year at the University of North Carolina,” a series of 44 weekly entries describing his activities as a University student. He graduated in 1842, received his MD from the University of Pennsylvania’s Medical Department in 1845, and returned to Lexington to practice medicine. During the Civil War, he served with the Fourteenth Battalion, Rowan County Home Guard. Though he survived the conflict, three brothers, two brothers-in-law, and a niece died during the war years. After the war Dusenberry resumed his medical practice in Lexington and served as a UNC trustee from 1874 until 1877. He died on 24 February 1886 and was buried in the Lexington City Cemetery. He never married.

Dusenberry's journal, the centerpiece of this new digital collection, provides multiple opportunities to

extend his text by creating a multimedia scholarly apparatus that, when combined with an array of interpretive essays, will illuminate the academic, social, political, economic, and religious forces that shaped his world. Though Dusenberry was not “famous” in the ways that our culture assigns such prominence, like many students today he enjoyed his senior year in college. He appreciated his friends; enjoyed sports, music, and dance; and despite an active social life, completed his studies successfully and spent his life as a physician in Lexington, North Carolina. The journal is a valuable source of information for those interested in antebellum culture, antebellum literary life, and the day-to-day events that ordinarily fall through the cracks of history. Edward L. Ayers, southern historian and one of the pioneers of digital libraries, points out that new forms of digitization and spatial display enable scholars and students alike to “see things that are invisible otherwise”.¹ The Dusenberry Journal's multimedia apparatus will allow users to both see and hear a slice of American history. All of the materials included on the site will be accompanied by scholarly annotations, biographies, and essays that will provide an analytical framework for the project and forge connections between the disparate materials (and disciplines) represented. When it is completed, The Dusenberry Journal will be a fully realized, searchable, multimedia, scholarly edition consisting of manuscript materials, images, songs, artifacts, maps, newspaper clippings, court and judicial documents, and important related resources pulled together from a variety of repositories, especially the University Library's special collections; the North Carolina State Archives; North Carolina public libraries; and the private collection of a family descendant, Colonel William B. Hankins, Jr. The scholarly apparatus for Dusenberry's journal will be accessible to users by means of links within the edited text and through various indexes for personal names, places, publications, images, topics, events, dates, organizations, genres, and authors.

2. Technology

Digital images of the pages of the journal have been captured, and the text has been marked up in TEI P5 XML. The handwritten text has been traced and output in a Scalable Vector Graphics (SVG) format. SVG is itself an XML format, which means structures (i.e. lines, words, and letters) in the image can be linked to lines and notes in the transcribed text. Open Source web mapping software (OpenLayers) is being used to provide zoomable overlays of the SVG and raster image for each page. The result is an interface in which each line of text in the transcription is linked to a line of written text on the page image. The page image and transcription are displayed side-by-side,

and OpenLayers provides zoom and pan features for the image.

The img2xml system models tracings of manuscript text as Shapes: the SVG paths and bounding boxes; Regions: bounded spaces containing text; and Structures: the overlap of one or more shapes with a Region. Structures can be mapped to elements in a transcription or to annotations.

Since SVG is an XML-based format, it can be manipulated in a web browser using standard Javascript techniques. The final project is available at <http://docsouth.unc.edu/dusenberry>

The digital environment has the power to contextualize and fully document this ordinary life, proving, as Nell Sigmon put it, "You don't have to be famous for your life to be history".² In that, we fully realize one of the most distinguishing features of electronic editions - "their capaciousness: scholars are no longer limited by what they can fit on a page or afford to produce within the economics of print publishing".³

References

Burnard, Lou, Bauman, Syb (eds.) (November 8, 2009). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. ver. 1.5.0. <http://www.teic.org/release/doc/tei-p5-doc/en/html/index.html>.

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

Cayless, Hugh (2009). 'Image as Markup: Adding Semantics to Manuscript Images'. *DH 2009*. University of Maryland, USA, 22-25 June 2009. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf.

Notes

1. The Chronicle of Higher Education, 10 November 2006: 33
2. Jacqueline Dowd Hall, interview with Nell Putnam Sigmon, 13 December 1979 (H-143), Southern Oral History Program Collection #4007, Southern Historical Collection, University of North Carolina at Chapel Hill.
3. Price, Kenneth (2008). "Electronic Scholarly Editions," in A Companion to Digital Literary Studies, ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell, 2008.

Delivering virtual reality: a proposal for facilitating pedagogical use of three-dimensional computer models of historic urban environments

Lisa M. Snyder

lms@ucla.edu

UCLA's Academic Technology Services/Institute for Digital Research and Education and the Urban Simulation Team at UCLA University of California, Los Angeles

Scott Friedman

friedman@ucla.edu

UCLA's Academic Technology Services/Institute for Digital Research and Education and the Urban Simulation Team at UCLA University of California, Los Angeles

Examination of the built environment is a fundamental line of humanistic inquiry that shapes our understanding of diverse cultures. It is impossible to consider the ancient Egyptians without immediately thinking of the pyramids or the vast religious complexes of the Nile river valley. The medieval pilgrimages and the modern disciplines that study them – history, literature, religion, musicology, and art – are inextricably tied to the monumental Romanesque cathedrals that blanket Europe. This link between humanities scholarship and the built environment is unquestionable, with an inexhaustible number of examples that illustrate how architecture and urban design reveal the aspirations and priorities of cultures across the ages.

The humanistic promise of virtual reality environments created to academically rigorous standards is that they would allow scholars the opportunity to explore reconstructed buildings and urban spaces, recreate the experience of citizens from other eras, gain insights into the material culture of past civilizations, and engage formal and informal learners in ways never before possible. Within these virtual spaces, students could navigate through reconstructions of historic urban environments and engage with one another to build knowledge through constructivist learning exercises. Beyond pedagogy, this new form of academic expression would engender new forms of scholarship and publication.

Seemingly, the promise of the technology is within reach. Scores of researchers across the globe are hard

at work on a significant – and growing – body of academically generated three-dimensional content. Architects and archaeologists have embraced three-dimensional computer modeling to visualize unbuilt structures,¹ reconstruct cities of the past,² and explore research questions specific to a single site or structure.³ Yet, even given all that promise and hype, two-dimensional drawings and static images still dominate the research, study, and teaching methods of architectural and urban form. The problem is that only a very small selection of digital work has been made available to scholars, students, and the general public. In part, this is because of a dearth of options for easily sharing three-dimensional content. Available on-line social-networking virtual worlds such as Second Life and Stanford's Sirikata, while brilliant for avatar interactions and communication, are not suited to the study of highly detailed environments. The platforms for most online virtual worlds do not support the ability to import or export content, thereby limiting the scope and functionality of the exploration to a single platform. The strength of available online mapping programs like Google Earth and Microsoft Research Maps is in their ability to let users interact with rich geographical content on a global scale, not their replication of the physical world and first person navigation. While Google actively encourages users to create three-dimensional content with its free modeling package (Sketch Up), there are extreme limitations to using Google Earth for the study of the built environment. Navigation is geared towards bird's-eye examination of schematic forms, and the models that can be loaded in Google Earth are, by necessity, very low resolution. When the goal is to illustrate the schematic massing of a city, these limitations are acceptable, but not when the goal is to understand the experience of moving through a detailed world. A secondary problem is that the computer models, by themselves, are essentially raw data, lacking contextual material, subject expert commentary, and textual analysis that could make them engaging and effective tools for teaching and learning.

The recent explosion of interest in online worlds, mapping software, and three-dimensional games is reinforcing both consumer interest in three-dimensional worlds and the pressing need for software to interrogate three-dimensional content created to rigorous academic standards. While there are many different types of software and online opportunities for interacting with three-dimensional computer models, nothing currently available addresses the unique requirements of humanities scholars and students. What is needed is a concerted effort to move these research projects successfully from the research lab to mainstream scholarship and pedagogy across the disciplines. **This presentation will discuss efforts at UCLA**

by Drs. Lisa M. Snyder and Scott Friedman to leverage existing and new modeling work for broad pedagogical use by creating a real-time software interface and content repository to provide a mechanism for exploring highly detailed three-dimensional models in educational settings.

The first phase of the project, currently underway, is the creation of generalized and extensible run-time software that will allow real-time exploration of highly detailed, three-dimensional computer models in both formal and informal educational settings. **This run-time software will address the greatest challenge for building knowledge through use of three-dimensional computer models by providing scholars and educators the mechanism to explore, annotate, craft narratives, and build arguments within the three-dimensional space** – in essence, facilitating the creation of virtual learning environments that can be broadly disseminated to educators and learners across grade levels and humanities disciplines. This software will also allow the raw computer models to be used as the basis for constructivist learning activities, so that students can actively engage with the content to build knowledge by creating a personalized virtual learning environment.

This effort will provide an innovative way for scholars, students, and the general public alike to interrogate academically generated three-dimensional content. Scholars will be able to expand their own knowledge through interactive exploration of the modeled environment, and, using the proposed authoring tools, develop their own arguments and class presentations within the virtual space. By capitalizing on student interest in three-dimensional content as evidenced by the popularity of online worlds and mapping programs, this new software will expose students to academically vetted content, encourage them to develop critical thinking skills about the historic reconstruction process, and provide an alternative to the traditional classroom methods for examining the built environment. Lifelong learners will also be able to access the content and build their own knowledge about past cultures. The availability of the proposed software will also create opportunities to leverage existing modeling projects by providing a mechanism for sharing content, and encourage new work that can take advantage of the new run-time software.

The educational promise of digital computer environments has yet to be realized, largely because past efforts have focused on the short-term constructivist benefits of the process for the academic development team. As a result, a great deal of content has been developed but is unavailable for general use. The proposed repository and administrative

front end that constitutes the second phase of the project will facilitate submission of this scholar-created content,⁴ allow the aggregation of multi-media support material, and provide incentives for contributors to share their data. In addition to providing them access to the open source software, this administrative front end would ensure that content contributors are given proper credit for their work, and provide them a mechanism to control how they distribute their content and charge for their work.⁵

1. Background

UCLA has a long history with three-dimensional computer modeling and on the development of real-time software for exploration of virtual environments. Given this long history, work on various fronts has been building to the development of this run-time software and content repository since 1996. The idea for an educational repository for real-time content was promoted in Snyder's 2003 dissertation and was based on discussions with Dr. Scott Friedman (UCLA Academic Technologies Services) in the preceding years. Work on computer reconstructions of historic urban environments has been conducted through the Urban Simulation Team at UCLA since 1996 and through the Experiential Technologies Center (the successor lab to the CVRLab) since 1997.⁶

2. The authors

Dr. Lisa M. Snyder has been a member of the Urban Simulation Team since 1996. She is also the associate director for outreach and operations for the UCLA Experiential Technologies Center which operates under UCLA's Academic Technology Services and the Institute for Digital Research and Education. Her research is focused on the educational use of interactive computer environments. Through the UST, she developed the real-time simulation models of the Herodian and Umayyad Temple Mount site now installed at the Davidson Center in Jerusalem (see Figures 8-19) and is currently working on the computer reconstruction of the World's Columbian Exposition of 1893 (see Figures 1-7).

Dr. Scott Friedman is a computer scientist with UCLA's Academic Technology Services and, since 1994, has been the principal developer of the Urban Simulation Team's software systems. Friedman participated in the NSF funded Virtual World Data Server project. His role in that project focused on integrating an Urban Simulator client into that system to support multiple users and very large data sets. He specializes in multimedia systems, interfaces for three-dimensional environments, and distributed

computing. His 2003 dissertation focused on "The Pixelcluster: Real-time visualization using a cluster of commodity workstations."

3. Funding

This work is being supported by UCLA's Academic Technology Services/Institute for Digital Research and Education and the National Endowment for the Humanities.





Figures 1-5. Screen snapshots taken from the real-time visual simulation model of the Herodian Temple Mount developed jointly by the Urban Simulation Team at UCLA and the Israel Antiquities Authority. From top to bottom: the Temple Mount; the north/south Roman road and Robinson's Arch; on the platform; in the Royal Stoa; a view of the Second Temple from within the Royal Stoa.

University of California, Los Angeles. <http://www.ust.ucla.edu/ustweb/ust.html> (accessed 12 March 2010).

Notes

1. Pre-visualization modeling and rendering is now commonplace at both the professional level and in architecture schools.
2. Computer modeling has been embraced by the Computer Applications and Quantitative Methods in Archaeology (CAA) community, with a broad range of projects that showcase computer modeling.
3. Research questions within the modeling environment might include viewshed analysis, reconstruction alternatives, and placement of statuary.
4. This type of content repository has precedent in the 'Great Buildings' series created in the mid-1990s. It was also a central element of Snyder's doctoral dissertation, and the proposal of Koller et al. to archive computer models of cultural heritage sites.
5. This is an important feature in that it would allow content providers the opportunity to generate a revenue stream to support ongoing development and help to make large-scale modeling projects self-sustaining.
6. Models constructed through the UCLA Experiential Technologies Center include the NEH-funded model of Karnak, John Dagenais' reconstruction model of the Romanesque Cathedral of Santiago de Compostela, Diane Favro's work on ancient Rome, new student work underway in Egypt under the direction of Willeke Wendrich (including models of Fayuum and Saqqara). Work through the Urban Simulation Team includes the Snyder's model of the World's Columbian Exposition of 1893 and reconstructions of the Temple Mount in Jerusalem in both the first and eighth centuries.

References

The Experiential Technologies Center (2005-2010). *The Experiential Technologies Center, University of California, Los Angeles.* <http://www.etc.ucla.edu/> (accessed 12 March 2010).

Friedman, S. (2004). The Pixelcluster: Real-time visualization using a cluster of commodity workstations. Dissertation University of California, Los Angeles.

Koller, D., Frischer, B., Humphreys, G. (2009). 'Research challenges for digital archives of 3D cultural heritage models'. *Journal of Computing and Cultural Heritage.* 2(3): article 7. <http://portal.acm.org/citation.cfm?id=1658346.1658347> (accessed 12 March 2010: via ACM Portal, UCLA).

Snyder, L. (2003). The design and use of experiential instructional technology for the teaching of architectural history in American undergraduate architecture programs. University of California, Los Angeles, Dissertation.

The Urban Simulation Team at UCLA (1997-2010). *The Urban Simulation Team,*

Digitizing Ephemera and Parsing an 1862 European Itinerary

Tomasek, Kathryn

ktomasek@wheatonma.edu

Wheaton College, Norton, Massachusetts

Stickney, Zephorene L.

zstickne@wheatonma.edu

Wheaton College, Norton, Massachusetts

This interactive digital poster demonstrates the advantages of digital publication over print for a particular kind of socio-historical project. It uses as an example three incomplete sets of sources: a travel journal, herbaria, and ephemera from an 1862 tour of England and Europe that was undertaken by Wheaton College founder Eliza Baylies Wheaton, her husband Laban Morey Wheaton, and his cousin David Emory Holman. Linking TEI-compliant XML text with images of these sources, our poster/demo offers an approximation of the experiences that led to the collection of the items. Our interactive digital poster also includes links with historical maps of England, Wales, France, Italy, Switzerland, the Rhine Valley, and Belgium. Clicking on a location brings up relevant sections of the interpretive historical essay as well as images of relevant pages of the travel journal, herbaria, and ephemera. An interactive timeline offers an alternate method of accessing the data.

During the 1862 journey, Eliza Baylies Wheaton kept receipts for her housekeeping transactions in London, and she compiled a travel journal and herbaria, thus leaving for the historian multiple genres of accounts of her interests and experiences — financial, descriptive, and botanical. The resulting narratives convey the texture of daily life for a nineteenth-century traveler and reflect the wide-ranging interests of a woman who cherished her husband and friends, loved art and gardens, practiced devout Christianity, painstakingly recorded the engineering details of the new tunnel under the Thames that connected London to Greenwich, and pursued every opportunity to visit sites associated with Napoleon Bonaparte. Such narratives are conveyed less than optimally in traditional print publications, at least partly because cost considerations would prohibit inclusion of full-color plates for presenting such an obscure collection. These texts and collections have been digitized as part of the Wheaton College Digital History Project.

Digital presentation allows interactive viewing of the document images that we suggest might approximate the series of experiences that led to collection of the ephemera and specimens for the herbaria alongside the recording of the travel journal. Further, including links to images of the primary sources introduces a kind of transparency that is missing from traditional print methods for presenting results of historical research. Digital presentation enhances the historian's ability to recreate a past that all too often remains obscure — a set of events from daily life that includes not only the experiences of well-to-do tourists who created and collected the items in archival collections but also the boardinghouse keepers, laundresses, and shopkeepers with whom they interacted. Digitally presented history can be social history at its best.

1. Financial Records

Beyond its local interest for friends and alumni of Wheaton College, the project has larger historical value in its attention to the 1862 journey in the context of changing economic conditions in Great Britain, Europe, and the United States in the mid-nineteenth century. The group of travelers who created the archive presented in this poster/demo combined business with tourism while in London, as the two men shared interest in the production of straw hats. Holman took with him on the journey a prototype that demonstrated his innovation for machines used to shape the crowns of hats, and he established residency in London to begin the process of registering a patent for his machine. A patent drawing has been found at the British Library. While Holman continued to board in London, Eliza Baylies Wheaton and her husband toured in the south and west of England and in the south of Wales, and they traveled in Europe for two and a half weeks in July 1862.

The journey also represented a transitional moment in the economic experiences of a well-to-do white woman from the United States. The poster/demo thus builds on and contributes to the growing historical literature about Anglo-American women and their economic lives in North America in the eighteenth and nineteenth centuries. Like many articles and monographs, this portion of the project focuses on the records left by an individual, digging down into the archival record to explore the financial details of a moment in one woman's life and explicate their larger historical significance.

The ephemera that the Wheatons collected included such seemingly mundane materials as laundry lists, boarding accounts, and receipts from restaurants and hotels. Such materials resemble the household accounts that Eliza B. Wheaton was accustomed

to keeping at home, and they demonstrate her continued responsibility for economic interactions with women workers while she and her husband and their traveling companion were away from home. Examining her household accounts alongside the social narrative she created in the travel journal demonstrates parallels between the pleasant tasks of sociability and the more quotidian concerns of housekeeping, whether Wheaton was at home or away. The herbaria add still another dimension, augmenting comments in the travel journal on such engineering feats as the Thames tunnel with botanical specimens identified according to the historical or cultural sites where they were collected to offer a view of the traveler as scientific collector, of both facts and specimens.

Since Eliza B. Wheaton was widowed three years after she and her husband returned to their home in Massachusetts, the financial records from the European journey document a significant moment in her economic life. They supplement a large number of cashbooks and other financial documents that she accumulated over the next forty years. Collected during a transitional period after she had begun to learn the details of her husband's business affairs, the 1862 receipts suggest the kinds of financial responsibilities to which Wheaton was accustomed during her marriage and the preparation that keeping household accounts gave her for handling her investments and managing her wealth after her husband's death. The travel journal and herbaria combine with other ephemera to document the interests that she shared with her beloved husband and the pleasures of their European adventure. The richness of this documentary collection and its multiple genres make digital publication the most appropriate method of dissemination.

2. Itineraries

In our exploration of the documents and ephemera that survive from the 1862 journey, we spent considerable time focused on our travelers' itineraries and how we know them, paying special attention to what the materials in the collection do and do not tell us. Eliza Baylies Wheaton's travel journal is incomplete and is only one of many items in the collection that document the Wheatons' European summer. We have used the collection's ephemera — which include trade cards, lists, and notes, in addition to the receipts, herbaria, and boarding accounts — to fill in gaps in the account of the journeys contained in the travel journal. Our poster/demo includes images of these documents and their XML coded transcriptions.

The travel journal stopped after a description of a journey to Windsor on May 29, 1862, and then picked

up again in mid-July when the Wheatons crossed the English Channel to the Continent. Eliza Baylies Wheaton's herbaria show us that during June, she and her husband traveled around southern England with London as their base. The herbaria have posed particular challenges for scanning, transcription, coding, and interpretation. Although none of the pages are dated, they reveal the couple's itineraries outside London and act as a nineteenth-century version of a photo album. The herbaria also raise questions about the sites the Wheatons actually visited; Italy was a particular puzzle.

The Wheaton Family Papers includes a copy of Samuel Rogers's *Italy: A Poem*, an example of popular culture in the United States connected to European tourism in the mid-nineteenth century. The book's presence in the collection suggests that Eliza Baylies Wheaton understood the importance of Italy as the focus of European tourism since the beginnings of the Grand Tour. Yet other than the herbaria, there are no documents from the journey to suggest our travelers' presence in Italy. Whilst the herbaria once led us to believe that the Wheatons visited Florence, Rome and Pompeii, our parsing of their itinerary through surviving hotel receipts precluded the possibility of their having had time to visit Italy during this trip. Perhaps most significantly as we sought to understand our travelers' motivations and actions, we considered the political instability of Italy in 1862, noting that Garibaldi's army marched on Rome in July and August. Herbaria pages regarding Rome include pasted-in images, apparently cut from a set of small photographs, probably because the Wheatons did not actually see these sites.

Our reading of the laundry lists and boarding house receipts combines with our parsing of the European itinerary to tell particular stories about the Wheatons' travels in England and Europe in the spring and summer of 1862. Our being able to display images and transcriptions of the primary documents alongside our interpretations gives audiences the opportunity to weigh our analysis and comment upon it in ways that are quite foreign to the usual methods of presenting historical analysis.

References

- Bendixen, Alfred, Hamera, Judith (eds.)** (2009). *Cambridge Companion to American Travel Writing*. New York: Cambridge University Press.
- Boydston, Jeanne** (1990). *Home and Work: Housework, Wages, and the Ideology of Labor in the Early Republic*. New York: Oxford University Press.

- Buzard, James** (1993). *The Beaten Track: European Tourism, Literature, and the Ways to Culture, 1800-1918*. Oxford: Clarendon Press.
- Cohen, Daniel J., Rosenzweig, Roy** (2006). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. Philadelphia: University of Pennsylvania Press.
- De la Peña, Carolyn, Vaidyanathan, Siva (eds.)** (2006). 'Special Issue: Rewiring the "Nation": The Place of Technology in American Studies'. *American Quarterly*. **58/3**: 555-985.
- Dublin, Thomas** (1994). *Transforming Women's Work: New England Lives in the Industrial Revolution*. Ithaca, N.Y.: Cornell University Press.
- Earhart, Amy**. 'Mapping Concord: Google Maps and the 19th-Century Concord Digital Archive'. *dhq: digital humanities quarterly*. <http://digitalhumanities.org/dhq/vol/3/3/000057/000057.html>.
- Elliott, Tom, and Sean Gillies**. 'Digital Geography and Classics'. *dhq: digital humanities quarterly*. <http://digitalhumanities.org/dhq/vol/3/1/000031/000031.html>.
- Gamber, Wendy** (1997). *The Female Economy: The Millinery and Dressmaking Trades, 1860-1930*. Urbana, Illinois: University of Illinois Press.
- Gamber, Wendy** (2007). *The Boardinghouse in Nineteenth-Century America*. Baltimore: The Johns Hopkins University Press.
- Ginzberg, Lori D.** (1990). *Women and the Work of Benevolence: Morality, Politics, and Class in the 19th-Century United States*. New Haven, Conn.: Yale University Press.
- Hartigan-O'Connor, Ellen** (2005). 'Abigail's Accounts: Economy and Affection in the Early Republic'. *Journal of Women's History*. **17/3**: 35-58.
- Helmreich, Paul C.** (2002). *Wheaton College, 1834-1957: A Massachusetts Family Affair*. New York: Cornwall Books.
- Jensen, Joan** (1986). *Loosening the Bonds: Mid-Atlantic Farm Women, 1750-1850*. New Haven, Conn.: Yale University Press.
- Kelley, Mary** (2006). *Learning to Stand and Speak: Women, Education, and Public Life in America's Republic*. Chapel Hill, N.C.: University of North Carolina Press.
- Kelly, Catherine** (1999). *In the New England Fashion: Reshaping Women's Lives in the Nineteenth Century*. Ithaca, N.Y.: Cornell University Press.
- Knowles, Anne Kelly, ed.** (2002). *Past Time, Past Place: GIS for History*. Redlands, California: ESRI Press.
- Knowles, Anne Kelly, ed.** (2008). *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*. Redlands, California: ESRI Press.
- Luton Public Museum** (1933). *The Romance of the Straw Hat, Being a History of the Industry and a Guide to the Collections*. Luton: Luton Public Museum.
- Matson, Cathy** (2006). 'Women's Economies in North America before 1820: Special Forum Introduction'. *Early American Studies*. **4/2**: 271-290.
- Millar, Marla R.** (2006). *The Needle's Eye: Women and Work in the Age of Revolution*. Amherst, Mass.: University of Massachusetts Press.
- Nash, Margaret** (2005). *Women's Education in the United States, 1780-1840*. New York: Palgrave Macmillan.
- Nead, Lynda** (2000). *Victorian Babylon: People, Streets, and Images in Nineteenth-Century London*. New Haven, Conn.: Yale University Press.
- Paine, Harriet E.** (1907). *The Life of Eliza Baylies Wheaton: A Chapter in the History of the Higher Education of Women*. Cambridge: Riverside.
- Rogers, Samuel** (1842). *Italy: A Poem*. London: Edward Moxon.
- Schreibman, Susan, Ray Siemens, and John Unsworth, eds.** (2004). *Companion to Digital Humanities*. Oxford: Blackwell.
- Schriber, Mary Suzanne** (1997). *Writing Home: American Women Abroad, 1830-1920*. Charlottesville: University of Virginia Press.
- Stansell, Christine** (1982). *City of Women: Sex and Class in New York, 1789-1860*. New York: Knopf.
- Thomas, William G., and Edward L. Ayers** (2003). 'The Differences Slavery Made: A Close Analysis of Two American Communities'. *American Historical Review*. **108/5**: 1299-1307. <http://www2.vcdh.virginia.edu/AHR/>.
- Ulrich, Laurel Thatcher** (1990). *A Midwife's Tale: The Life of Martha Ballard, Based on Her Diary*. New York: Vintage.
- Wulf, Karin** (2000). *Not All Wives: Women of Colonial Philadelphia*. Ithaca, New York: Cornell University Press.

Critical Editing of Music in the Digital Medium: an Experiment in MEI

Viglianti, Raffaele

raffaele.viglianti@kcl.ac.uk

King's College London

Critical editions of music have not received the level of attention that research in Digital Humanities has given to textual criticism, which already has an established scholarly production of written contributions and digital publications. Digital representations in literary criticism are used for analytical purposes as well as for accommodating critical editions in the digital medium, which offers a high degree of interactivity and opens toward experimentation with new formats of publication. Nonetheless, there has been little debate about music editing in the new medium and only a few digital publications have been developed.

Several aspects of digital textual criticism can find an application on music documents because similar issues exist in the representation of primary sources and editorial intervention. In fact, since the early stages of music scholarship, musicologists looked at the editorial practices of classical philologists while working towards a definition of their own scientific finalities (Grier, 1996). The study of documentary sources transmitting a written work (manuscript or printed) is the main correlation between music and textual criticism. For instance, the discrepancy identified by Tanselle (1989) between text and the artefacts that transmit it, is a condition that applies to music notation as well as literature; however, literature is 'a one-stage and music a two-stage art' (Goodman, 1976:114; also discussed by Feder, 1987). A musical work, in fact, does not exist only as written notation but also requires performance to reach its final receiver. For this reason, understanding the complexity of the music work-concept and its associated cultural practices is central to the digital representation of music critical editions. Even though the recent research in digital textual scholarship provides a rich paradigm for the emergent field of digital editing of music, there is the need for more research on digital representation and publication of detailed notation data.

The work conducted for a postgraduate dissertation (MA) at King's College London attempted to discuss some of these issues. This poster presents the results of the dissertation's case study: a digital edition of Claude Debussy's *Syrinx* (*La Flûte de*

Pan) for flute solo. The XML-based model represents notation, variant readings and editorial intervention; additionally, several different views are extracted and rendered for presentation with vector images.

1. An experiment with MEI: *Syrinx (La Flûte de Pan)* by Claude Debussy

Syrinx (La Flûte de Pan) by Claude Debussy (1862 – 1918) is a short piece for flute solo originally composed as theatrical interlude for the play *Psyché* (1913) by Gabriel Mourey under the title *La Flûte de Pan*. Despite Debussy showing little interest in the publication of the piece, the first performer, Louis Fleury, contributed to the reception of the piece as independent from Mourey's play. The piece maintains a relevant role in the solo flute repertoire. Two principal sources have been used for the digital edition: the first edition published posthumously by Jobert in 1927 under the new title *Syrinx* and a recently discovered manuscript in a private collection in Brussels dated 1913 (MSB), which constituted the base text.

For the creation of a digital model for the new edition, the use of a combination of TEI and MusicXML was initially considered;¹ however, MusicXML does not match TEI's flexibility when encoding primary sources and variant readings. MusicXML, in fact, is primarily designed 'to be sufficient, not optimal' (Good, 2001); therefore, it represents normalised common western music notation to facilitate interchange and does not allow the flexibility in the granularity that is required when representing the editor's interpretation and understanding.

The Music Encoding Initiative (MEI) provided an alternative choice.² This XML format is modelled on TEI and attempts to follow the same principles. In particular it specifically focuses on formalising interpretation through declarative knowledge and claims to be independent from rendering software while also addressing processing matters (Roland, 2002).³ Moreover, it includes a module for the representation of variant readings and transcription of primary sources; therefore a combination with TEI did not seem to be necessary. The MEI format is still in development; however, the University of Paderborn (Germany) and the University of Virginia (USA) recently received a grant to take MEI out of its beta phase and produce public guidelines, which should be completed by the end of July 2010.⁴

2. The encoding model

The poster will show how the MEI model represents some editorial aspects common with textual editing (i.e. bibliographical metadata, correction and regularisation) and editorial issues related to the nature of the music notation (i.e. apparatus, rhythmic constraints, performative instructions). In particular it will show:

1. *The header*: similarly to TEI, MEI provides a “header” (<meihead>) that allows documenting information about the digital file and its sources. Notably, the elements in the description of the manuscript source MSB attempt to emulate the much more detailed encoding model offered by the TEI manuscript description module. The header also documents the encoding criteria for notation.

```
<measure>
  <staff>
    <layer>
      //-- Notes
      <note>, <beam>, <tuplet>, etc.
      ...
    </layer>
    //-- Phrase marks
    <slur>
      ...
      //-- Dynamics, tempo markings and
      directions positioned above the staff
      <dir>, <dynam>, <breath>, etc.
      ...
      //-- Dynamics, tempo markings and
      directions positioned below the
      staff
      <dir>, <dynam>, <breath>,
      etc.
      ...
    </slur>
  </staff>
</measure>
```

Example 1: Encoding criteria for notation

2. *Variant readings*: The MEI file represents this edition’s base text (MSB) and adds additional information every time a difference in the other sources occurs. If the sources agree, it is expressed silently. This criterion is identified by the TEI guidelines as ‘internal parallel segmentation’.⁵ Example 2 shows alternative notations encoded with <app> and <rdg>; the attribute type specifies which reading has been selected for the edition. It is worth explaining the basic mechanisms behind the element <slur>, since they are also common to other elements. The attribute staff defines to which staff the phrase mark belongs to; place defines whether the slur has to be rendered above or below the staff; tstamp identifies the beat in which the slur starts and dur the beat in which the slur ends.

```
<measure id="m28" n="28">
  ...
  <app>
```

```
<rdg source="MSB">
  <slur staff="1" tstamp="2" dur="4"
  place="above" />
</rdg>
<rdg source="FEJ" type="ed">
  <slur staff="1" tstamp="2"
  dur="2.875" place="above" />
  <slur staff="1" tstamp="3"
  dur="3.75" place="above" />
</rdg>
</app>
...
</measure>
```

Example 2: Alternative phrase marks in bar 28 from source MSB and FEJ.

3. *Editorial conjecture and intervention*: change of hand, additions, corrections and supplied notation have been encoded with elements equivalent to the ones employed by the TEI.
4. *Problematic cases*: bar 22 in MSB presents an incongruent rhythm and a missing barline to separate it from bar 23. The encoding used an empty-element version of <beam> to encode the differences in beaming resulting from the different rhythms and a <gap> element for the missing barlines.

3. Presentation

For this edition, a number of different perspectives from the MEI document are produced using XSLT 2.0 to serialize into a text format for Mup, a program that converts its own notation format into Post Script vector graphics (PS). To extract musical information from the encoded edition, a heavily customised version of a MEI to Mup XSLT provided on the MEI website under the Educational Community Licence. The poster will show the following views:

1. *The edited piece*. The XSLT script extracts all the variant readings chosen for the edition, includes the editorial interventions marked with <supplied> and creates a new MEI document tree containing only the notation for the edition. This tree is then serialized into Mup language and rendered in PS.
2. *Synoptic apparatus*. Instead of printing a traditional apparatus, this edition proposes a synoptic apparatus of the two main sources (MSB and FEJ). This is automatically built to display measures that contain variant readings. Moreover, it has been programmed to display the notation of the two sources in a semi-diplomatic manner and excluding editorial intervention.



Figure 1: synoptic display of variants in bar 32: Accent in FEJ but *decrecendo* in MSB

3. *Breath marks.* MSB has fewer breath marks than the first edition FEJ and it is possible that Debussy did not provide all the breath marks necessary for performance. Trevor Wye (1994), in his edition of *Syrinx*, introduces a number of recommended breath marks to support the performance. Since Wye employs MSB as a base text, his breath marks are suitable for this edition's notation as well. The encoding for this case study edition includes Wye's additions with the element <add resp="Wye">. The XSLT programmed to transform the edited music notation can, if requested, include these marks in the Mup output.

4. Future work

The views created for this prototype are static; however it would be highly desirable to combine them in an interactive environment. For example, the apparatus could be enhanced allowing moving measures on the screen to be compared. Performers might be interested in knowing different tempi and breath marks from other editions, like Wye's breath marks, and include them in the edition to be printed out with a printing device at home. Paper editions with a similar approach are not uncommon, but often include comparative tables of tempi and resolution of ornaments that cannot be directly included in the edited text (see Palmer, 1991).

Even though the Web is currently the preferred digital publishing environment, there still is not a straightforward method to output notation as HTML and possibly there will never be. The only possibility to publish music on the Web is through graphic information; however, systems for delivering complex interactivity based on image formats are becoming increasingly common. OpenLayers, for example, shows how images can be made highly interactive through the superimposition of layers.⁶ Future work will focus on implementing interactivity on the PS views in OpenLayers.

N.B. All websites mentioned in footnotes have been accessed 6 March 2010.

References

Bach, Johann S. (1991). *Inventions and Sinfonias (Two- and Three-Part Inventions)*. Palmer, W. A. (ed.). Van Nuys (CA, USA): Alfred Publishing.

Debussy, Claude (1927). *Syrinx*. Paris: Jobert.

Debussy, Claude (1991). *La Flûte de Pan*. Ljungar-Chapelon, A. (ed.). Malmö: Autographus Musicus, Facsimile.

Debussy, Claude (1996). *Syrinx (La Flûte de Pan)*. Stegemann, M., Ljungar-Chapelon, A. (eds.). Wien: Wiener Urtext Edition.

Feder, Georg (1987). *Musikphilologie: eine Einführung in die musikalische Textkritik, Hermeneutik und Editionstechnik*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Grier, James (1996). *The Critical Editing of Music: History, Method, and Practice*. Cambridge: Cambridge University Press.

Good, M. (2001). 'MusicXML for Notation and Analysis'. *Computing in Musicology*. 12: 113-124.

Goodman, Nelson (1976). *Languages of art: an approach to a theory of symbols*. Chicago: University of Chicago Press.

Roland, P. (2002). *The Music Encoding Initiative (MEI)*. <http://www2.lib.virginia.edu/innovation/mei/Papers/maxpaper.pdf> (accessed 6 March 2010).

Sperberg-McQueen, C. M. (2009). 'How to teach your edition how to swim'. *Literary & Linguistic Computing*. 24.1.

Tanselle, G. Thomas (1989). *A rationale of textual criticism*. Philadelphia: University of Pennsylvania Press.

Notes

1. MusicXML (distributed with a royalty-free license) is an interchange format developed by Recordare LLC (<http://www.recordare.com>). It is employed by free and commercial music editors (i.e. Finale and Sibelius) as intermediate representation to exchange data (often with loss).

2. <http://www2.lib.virginia.edu/innovation/mei/>

3. This approach recalls the claim of Sperberg-McQueen (1997) that 'declarative' information about the editorial process can be represented and modelled through markup: 'there is an infinite set of facts related to the work being edited' and 'any edition records a *selection* from the observable and the recoverable portions of this infinite set of facts'. This *selection* represents what the edition 'knows [...] about the work edited, its genesis and/or transmission, etc.'. This body of information is 'declarative'.

4. <http://www2.lib.virginia.edu/press/music/index.html>

5. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPPS>

6. <http://openlayers.org/>

LogiLogi: The Quest for Critical Mass

Wiersma, Wybo

mail@wybowiersma.net
King's College London

In this abstract, and more so in the poster-presentation, we will report on the process of, and the problems involved in, gaining a critical mass of users for an interactive hypertext application for the Digital Humanities. The aim of any DH application ultimately is to be used, but for collaborative ones, the contributions and interactions of existing users are often what make it worthwhile for new visitors. Gaining an initial critical mass of users for such applications is notoriously hard, but especially important if they are ever to be used at all.

First we briefly introduce LogiLogi, the system on which we are going to try to get a community started. Next our strategy for gaining users, some possible improvements, and attempts so far, are explained. Here we will also discuss the kinds of users we target, and the possible size of the application's critical mass. We finish with an overview of the usage-data that our poster will report on.

1. System

LogiLogi is a Web 2.0 application that tries to find an informal middle-road between good conversations and journal-papers by providing a form of quick, informal publication, peer-review, and annotation of short philosophical texts. It is intended for all those ideas that one cannot turn into a full-sized paper, but that one deems too interesting to leave to the winds.

It does not make use of forum-threads (avoiding their many problems), but of tags and links that can also be attached to phrases by people other than the text's original author. It also features a rating-system modelled after journal-based review. Well-rated texts earn authors more voting-power, and thus a measure of standing, within their peer group (of which there are multiple).

LogiLogi is Free Software, and has been under development for 3 years, during which about 30 volunteers have done 8 man-years of work (worth \$500.000). A public beta is already online and fully functional at <http://www.LogiLogi.org>

2. Strategy

Things that have been done so far to gain users are, first of all, making sure that LogiLogi works properly. LogiLogi has been extensively tested and improved at the LIRMM lab of the University of Montpellier this September. And it was used there by about 30 active users for internal discussions until the end of October. Secondly, some seed-content has been added (about 100 philosophical texts, some of which are part of larger essays). And finally, since October, it has been made easy for users to track new replies, annotations, and votes for their documents, both through a personalized RSS feed, and e-mail alerts. These things have made LogiLogi practically usable for the first time.

2.1. Target Audience

LogiLogi has not yet been advertised widely, and changing this is one of the first things we will do next. LogiLogi aims for a wide audience of scholars, students, and people interested in philosophy, but to set the right tone, we first aim for people with academic credentials (students and scholars). Among them, most success is expected with students, both because of their limited access to other publishing channels, and their greater average computer-literacy. Possible places to reach them are forums, newsgroups, and (limited) advertising via Google Adwords.

2.2. Process

Then, as part of user-driven, agile development, feedback will be collected from users on possible improvements: both ongoing, from users on the web, and from a small group of philosophers/students in a usability test. Some of these improvements will then be implemented, after which we plan to repeat the process, with another round of usability testing and improvements.

2.3. Improvements

A possible improvement so far identified is simplifying the application, for example by (temporarily) limiting the number of voting-communities (peergroups) to one. This would have the additional advantage of reducing the number of users that are needed to reach critical mass, because votes are no longer limited to, and divided between groups. While it is hard to determine what the critical mass of LogiLogi would be, from what we saw in the LIRMM case, it most likely lies between 30 to 60 active users per peergroup (or for the whole site, if there is only one peergroup). To examine this further, a small literature study of the notion of

critical mass, and of the factors influencing its size (especially for hypertext based applications, close to the humanities) will also be undertaken.

Another place for improvement is the editing and annotation process: its responsiveness, especially, could be improved. LogiLogi currently requires people to open a new page when they want to insert annotations or links, while it would be a lot easier if this could be done while reading the text, at least for simple annotations. And finally, a demo-video will be created, which quickly explains what LogiLogi is, and how it can be used.

2.4. Report

In our poster we will present LogiLogi, explain the notion of critical mass, and report on developments in the number of users. In addition, the strategies and improvements we applied, and their practical, and causal relationships will be explained, where possible. Also, we will not just be reporting the number of registered users, or unique visitors, but also on the number of documents, annotations, replies, and votes given over the time-period from December 2009 until June 2010. Thus a detailed view will be given of the process of gaining critical mass.

3. Conclusion

Whether we succeed or not in gaining a critical mass for LogiLogi, there will be meaningful results from this experiment, as it not only involves presenting, or further improving an already quite usable interactive Digital Humanities application, but foremostly trying to give it a critical mass of users, and exploring this process, producing insights and a valuable case-study (of success or failure) for future Digital Humanities projects to learn from: projects which will, most likely, be more interactive than their predecessors, and thus will sooner or later face the same challenge of gaining a critical mass of users.

Acknowledgements

We are grateful to Lars Buitinck, Maarten Geraedts, Allan van Hulst, Auke Klazema, Bart Leusink, Miguel Lezama, Charl Linssen, Jan Mikac, Steffen Michels, Roel 3 van Rijswijk, Bruno Sarlo, Thierry Stamper, Artyom Syazantsev, Rens van Summeren, Pieter van der Vlis, Jordy Voesten, Ilona Wilmont, Andrew Wolters and Feng Zhu for their contributions to the development of LogiLogi over the years. Among them we want to especially thank Bruno, Charl, Miguel, and Steffen, without whom LogiLogi would not have been what it is today.

We would also like to thank the Philosophy Department of the University of Groningen for the initial small grant that got LogiLogi started, and

the University of Nijmegen which twice provided a group of Computer and Information Science students to work on LogiLogi for credits as part of their GIPHouse program. In addition we are grateful to the OFSET (Organization for Free Software in Education and Teaching) foundation for supporting us with a small grant.

We would also like to thank the audiences of our presentations at the FOSDEM (Free and Open source Software Developers' European Meeting) of 2007 and 2009 in Brussels, the TDOSE (Technical Dutch Open Source Event) of 2007 in Eindhoven, the Netherlands, the ECAP (European conference on Computing and Philosophy) conferences of 2008 and 2009 in Montpellier and Barcelona, the Digital Humanities 2008 conference in Oulu, Finland, the RMLL (Rencontres Mondiales du Logiciel Libre) of 2008 in Mont-de-Marsan, France, the FKFT (Free Knowledge, Free Technology Conference) of 2008 in Barcelona, and the Philosophers Rally of 2009 in Enschede, the Netherlands, for their questions and insightful comments.

References

- Abrahamsson, P. et al.** (2002). 'Agile Software Development Methods: Review and Analysis'. *VTT Publications*. **478**.
- Baez, M. and Casati, F.** (2009). 'Liquid Journals: Knowledge Dissemination in the Web Era'. *LiquidPub Site*.
- Ball, Philip** (2005). *Critical Mass: How One Thing Leads to Another*. London: Arrow Books Ltd.
- Economides, Nicholas and Himmelberg, Charles** (1995). *Critical Mass and Network Size with Application to the US Fax Market*. Tech. rep. New York: New York University, Leonard N. Stern School of Business, Department of Economics. <http://ideas.repec.org/p/ste/nystbu/95-11.html>.
- Evans, Mark** (2010). 'What Makes a New Service Sticky?'. <http://www.markevanstech.com/2007/08/03/what-makes-a-new-service-sticky/>.
- Gladwell, Malcolm** (2002). *The Tipping Point: How Little Things Can Make a Big Difference*. Abacus, New Edition.
- Grafton, Anthony** (2009). *Worlds Made by Words: Scholarship and Community in the Modern West*. Harvard: Harvard University Press, 1st ed.
- Katz, Michael L. and Shapiro, Carl** (1985). 'Network Externalities, Competition, and Compatibility'. *American Economic Review*. **75**.3:

- 424-40. <http://ideas.repec.org/a/aea/aecrev/v75y1985i3p424-40.html>.
- Kemper, Andreas** (2009). *Valuation of Network Effects in Software Markets: A Complex Networks Approach*. Heidelberg: Physica-Verlag Heidelberg.
- Kolb, David** (2005). 'Association and Argument: Hypertext in and around the Writing Process'. *The New Review of Hypermedia and Multimedia: Applications and Research*. 1: 7-26.
- Kumar, V.** (2008). 'Critical Mass in E-Education'. *Eighth IEEE International Conference on Advanced Learning Technologies, 2008*. Santander, Cantabria, Spain, 2008, pp. 1009-1010.
- Levinson, Paul** (2001). *Digital McLuhan: A Guide to the Information Millennium*. Routledge, New Edition.
- Liebowitz, S. J.** (2010). *Network Externalities*. <http://www.utdallas.edu/~liebowit/palgrave/network.html>.
- LogiLogi.org** (2009). *Philosophy Beyond the Book*. <http://en.logilogi.org>.
- Marwell, Gerald and Oliver, Pamela** (2007). *The Critical Mass in Collective Action*. Cambridge University Press, 1st ed.
- Nadeau, Richard, Cloutier, Edouard and Guay, J. H.** (1993). 'New Evidence About the Existence of a Bandwagon Effect in the Opinion Formation Process'. *International Political Science Review/ Revue internationale de science pol.* 14.2: 203-213. <http://ips.sagepub.com/cgi/content/abstract/14/2/203>.
- O'Hear, Stephen** (2004). 'Critical mass'. *Times Educational Supplement*. 4564: 74.
- Papadimoulis, Alex** (2010). *The Great Pyramid of Agile - The Daily WTF*. <http://thedailywtf.com/Articles/The-Great-Pyramid-of-Agile.aspx>.
- Prasarnphanich, P. and Wagner, C.** (2008). 'Creating Critical Mass in Collaboration Systems: Insights from Wikipedia'. *2nd IEEE International Conference on Digital Ecosystems and Technologies, 2008*. Phitsanulok, Thailand, 2008, pp. 126-130.
- Prasarnphanich, P. and Wagner, C.** (2008). 'Explaining the Sustainability of Digital Ecosystems based on the Wiki Model through Critical Mass Theory'.
- Schachter, Joshua** (2010). *Del.icio.us: How tags exploit the self-interest of individuals to organize the Web for everyone*. <http://www.technologyreview.com/tr35/Profile.aspx?Cand=T&TRID=432>.
- Sundararajan, Arun** (2009). *Network Effects*. <http://oz.stern.nyu.edu/io/network.html>.
- The LogiLogi Foundation** (2009). *Software Libre for Your Web of Free Deliberation*. <http://foundation.logilogi.org/>.
- Wagner, C. et al.** (20XX). 'Creating a Successful Professional Virtual Community: A Sustainable Digital Ecosystem for Idea Sharing'.
- Wiersma, W. and David, S.** (2009). 'Two Scholarly Web-Agoras: The LogiLogi and Talia/Philospace Approaches'. *ECAP 2009 Abstract*. Barcelona, Spain, 2009.
- Wiersma, W. and Lezama, M.** (2008). 'LogiLogi: Combining Openness and Quality of Content'. *FKFT 2008 Proceedings*. 2008.
- Wiersma, W. and Sarlo, B.** (2008). 'LogiLogi: A Webplatform for Philosophers'. *Digital Humanities 2008 Book of Abstracts*. Oulu, Finland: University of Oulu, pp. 221-222.
- Wiersma, Wybo** (2010). *LogiLogi: Philosophy Beyond the Book*. <http://www.logilogi.org/pub/beyond/paper.pdf>.

Software Demonstration, “Emergent Time” timeline tool

York, Christopher

cyork@mit.edu

HyperStudio, Massachusetts Institute of Technology

Trettien, Whitney

HyperStudio, Massachusetts Institute of Technology

Emergent Time is a prototype collaboration tool for humanists and social scientists working with timelines—narrative arrangements of events. In Emergent Time, timelines are owned by particular users, and represent the user's interpretive reading of a series of events. While an individual timeline “belongs” to a user, many of the events it interprets may be shared by other users and interpreted differently in their timelines. Users construct timelines individually, using a single form to build on events others created before them, or to create new events from scratch. The application thus balances personal expression and argument (in the form of individual timelines) with collaboration and shared work (in the form of raw events).

Throughout the prototype, clicking on an event in a timeline will show how other users have interpreted that particular incident. Thus, one can read horizontally to follow the argument of a given timeline, or depth-wise to jump between different timelines that interpret the same event from different perspectives.

1. Overview timelines

The prototype's salient feature is a set of overview timelines, built by analyzing the network of links between timelines and events within the community. These links indicate the most important events for a given topic. For example, a search for “John F Kennedy” might show the most highly-cited events in his life: birth, election, and assassination. To accomplish this, the prototype uses a proprietary implementation of Page & Brin's PageRank algorithm. Events that are linked to in many timelines are likely to be important to the community, and receive a high rank; conversely, timelines that interpret many important events receive a boost in rank. Emergent Time uses these ranks to indicate which event entries are regarded as most authoritative by the general populace of users, and displays them when given a matching topic.

2. Collaboration strategy

In Emergent Time only the author of a given event can revise it, but the community at large can add source critique comments and propose alternate versions of the event. The design intention was to spark general discussion about whether a given event's description is well-supported by the primary sources cited. Because many versions of a given event may exist, this encourages users to link to the version that is factually best-supported in their own timelines, while passing over those with poor evidentiary support or badly-formulated descriptions. Hence, using an event in one's own timeline constitutes both a signal of interest in the historical incident and a vote of confidence in the event author's scholarship. The collaboration workflow thus serves as a macrocosm of the scholarly publication process, allowing authors and readers to evaluate the evidence in support of a given interpretation, and to “vote with their feet” by citing it rather than another in their own work.

As a result the overview timelines will come to reflect not only which events are most important for the interpretive community, but also which versions of a particular event are most authoritative. This allows overview timelines to present the most influential event entries for a given topic, and to accommodate shifts in communal knowledge as new evidence is found and new interpretations of a given incident become normative.

3. In contrast to other tools

This collaborative strategy is intended to capture established conventions for historical analysis and source critique, and use the resulting citation networks to construct overview timelines that accurately reflect the community's current normative views. By distributing small bits of knowledge among many event entries, promoting general discussion of the veracity of each, and then allowing users to “vote” for a given version of the facts by including the event in their timeline, it addresses shortcomings in other collaborative digital humanities approaches:

- *Open-revision wiki.* An open wiki implements what might be called “last man standing” collaboration. The last person to edit an article has license to revise and amend all the others' work, potentially reshaping it to his own ends. Of course, wiki history allows others to revise it back, but this encourages “squatting,” or continually monitoring an article in order to control its contents.

- *Moderated wiki.* Some wikis establish an editorial bureaucracy to address these issues. However, this in effect defers interpretation to an appointed “expert,” much after the fashion of a

traditional encyclopedia (with the proviso that the general public can submit material for editorial consideration).

- *Voting systems.* Finally, simple voting systems that ask users to “rate this article” suffer from known problems with blind polling. Anonymity encourages arbitrary voting; users might vote multiple times or use incomparable rankings; and the population of elective voters is self-selecting. By contrast, Emergent Time’s collaboration model is designed to circumvent such problems, since users “vote with their feet” by citing one formulation of an event rather than another in their timelines, and no one user can dominate interpretation by being the last to revise. While this model is relatively new to digital collaboration tools, it is quite similar to traditional humanities footnote and endnote citations. It clearly marks authorship and source material for a given interpretation, encourages communal discussion of the adequacy of an author’s evidence, and holds authors accountable for their votes by embedding the citations within their work.

Even with a sparse demo data set, it is clear that the Emergent Time prototype achieves a successful balance between individual work (seen when viewing a particular timeline) and community connections (via the interpretation comparison popup, and the related timelines and related users links). It encourages users to focus on developing their own ideas, while still suggesting points of contact with the wide community — for example, in the event editor, which shows possible base events as the user enters information. The opening page’s list of recent community activity is well-suited to draw users into other work and give a sense of liveliness. Most importantly, even for small data sets, it’s clear that the overview timelines do actually reflect the community’s current notion of the “most important” events.

Putting Edmonton on the (Google) Map

Zwicker, Heather

hwzicker@ualberta.ca

University of Alberta, Canada

Engel, Maureen

mengel@ualberta.ca

University of Alberta, Canada

In this poster/demo, we will describe and analyze the experience of teaching English 486, “Producing the City.” An experimental course co-taught between Dr Heather Zwicker, Associate Professor of English, and Dr Maureen Engel, E-Learning Manager for the Faculty of Arts at the University of Alberta, English 486 is a hands-on, theoretically grounded capstone course in multimedia installations that takes the city of Edmonton, Canada as inspiration and object. Based on principles of collaboration and student-centered learning, the course takes the city as its primary text. Grounded in short Edmonton narratives and a range of urban theory, the course listened to the city, looked at the city, moved through the city, and explored the meanings of home. The sensory experiences of sound, sight, and movement were translated through student projects using digital photography, simple mapping, soundscapes, and video. Each of these assignments served as a scaffolding exercise to prepare students for the final collaborative project: a KML-authored installation designed for Google Earth.

This course did not take GIS as its object of study; rather, it took the city as its object and asked students to use various multimedia tools to express their critical and creative engagement with that city and its narratives. Various assignments asked the students to demonstrate and explore their learning through digital tools, not to engage with and analyse the potential of the tools themselves. We asked them to learn new ways of expressing their ideas, and to discover the affordances of digital technologies to their critical apparatus. The course raised multiple questions about discipline, pedagogy, theory, and technology. Our presentation offers a critical commentary on our successes and shortcomings, and demonstrates the importance – and surprising payoffs – of doing this sort of work with undergraduate students in the traditionally low-tech field of English literature.

Our poster presentation / demo will have three components. The first is an overview of the course, describing its intellectual aims and technical models.

We explore some concepts that are often taken for granted: what is a map, how does it organize information, how does the concept of "space" translate into "place"? We overview digitally mapped urban literature in sites like Imagining Toronto, City of Memory (New York), Hitotoki (Tokyo, Shanghai, Paris, Sofia), Concrete Dialogues (Perth Australia), and Artangel (London), as well as acoustic ecology sites like the London Sound Survey, the Montreal and New York Sound Maps, as well as the Open Street Map project. Sites like these open up both the concept of mapping and the conventions of narrative in interesting ways, playing with the synchronicity of the traditional map and the linearity of conventional narrative. And yet for reasons to do with Edmonton's scarce representations and relative youth, such models could not be translated wholesale into our course.

The second section of our poster presentation turns to the pedagogical implications of team-teaching digital media under the aegis of the English Department. Working with digital media requires both instructors and students to shift their expectations. Whereas the pedagogy we're familiar with frequently measures student learning by verbal articulation, whether oral or written, we instructors had to learn to put "discovery learning" to work in classrooms by letting students explore on their own, to a certain extent. The biggest challenge for the students in English 486 was not the technology per se, but rather the nature of the assignments. Instead of sole-authored papers, for instance, students had to learn to work collaboratively on sustained projects over the course of the semester. The course also demanded unfamiliar ways of reading and writing, in addition to mastering the specific digital tools. Students had to figure out which rhetorical techniques are transferable and adaptable to the digital realm, and which are not. They had to exercise critical skills on the visual culture ubiquitous to their personal, if not their academic, experience. Our presentation pays particular attention to the ways in which the students surprised and surpassed our expectations, and the key lessons both they and we learned from the shift in genre from the research essay to the digital story/argument. Key to this aspect of our presentation will be demos of actual student projects.

Part three will offer a critical analysis of the digital tools we used for representing Edmonton. In particular, we evaluate Google Maps and Google Earth as a technical platform for this kind of pedagogical work. Publicly available and free of charge, Google Maps and Google Earth have much to recommend them; they present a low barrier to entry, both financially and technologically. At the level of politics, relying on Google for fundamental

courseware is problematic – asking our students to expose their work to a massive commercial enterprise based in a foreign country was a difficult decision to make. At the level of pedagogy, any digital application will present students with specific narrative constraints – Google Earth can only *do* what Google Earth can do, and would that be sufficient for the task we set for our students? We will assess the extent to which these tools are enabling or limiting, particularly to students crossing genres from traditional academic prose. The hypercities tool (<http://www.hypercities.com>) was evolving in beta as the course progressed, and though it ultimately would have served our pedagogical goals more satisfactorily, reliability and practicality carried the day. That Google's ubiquity and stability were significant determinants in our pedagogical practice is instructive, if disheartening.

Looking back, the course demanded much of both its instructors and its students. The quality of the work the students produced, however, and the extent of their learning proved that using digital tools pushed students to go farther than conventional tools could have.

References

- Abrams, Janet, Hall, Peter (eds.)** (2005). *Else/Where: Mapping New Cartographies of Networks and Territories*. Minneapolis: University of Minnesota Press.
- Artangel*. <http://nighthaunts.org.uk>.
- Benjamin, Walter** (1978). 'A Berlin Chronicle'. *Reflections: Essays, Aphorisms, Autobiographical Writings*. Demetz, Peter, Jephcott, Edmund (eds.). New York: Schocken Books, pp. 3-58.
- Best of Open Street Map*. <http://bestofosm.org>.
- Debord, Guy** (1981). *Theory of the Derive*. <http://www.bopsecrets.org/SI/2.derive.htm>.
- de Certeau, Michel** (1984). 'Walking in the City'. *The Practice of Everyday Life*. Berkeley, CA: University of California Press, pp. 91-110.
- City of Edmonton** (2004). *Naming Edmonton from Ada to Zoie*. Edmonton: University of Alberta Press.
- City of Edmonton maps*. <http://maps.edmonton.ca>.
- City of Memory*. <http://www.cityofmemory.org>.
- Concrete Dialogues*. <http://dialogues.concrete.org.au>.

- Edmonton Geological Society** (1993). *Edmonton Beneath our Feet: A Guide to the Geology of the Edmonton Region*. Edmonton: Edmonton Geological Society.
- Edmonton Police Service crime map*. <http://crimemapping.edmontonpolice.ca>.
- Edmonton Stories*. <http://www.edmontonstories.ca>.
- Goyette, Linda** (2004). *Edmonton in Our Own Words*. Edmonton: University of Alberta Press.
- Gregory, Ian, Ell, Paul** (2006). *Historical GIS: Technologies, Methodologies and Scholarship*. Cambridge: Cambridge University Press.
- Harmon, Katherine** (2003). *You Are Here: Personal Geographies and Other Maps of the Imagination*. Princeton: Princeton Architectural Press.
- Healthy City*. <http://healthycity.org>.
- Hitotoki*. <http://hitotoki.org>.
- Horton, Marc, Sass, Bill** (2003). *Voice of a City: The Edmonton Journal's First Century 1903 to 2003*. Collum, Peter (ed.). Edmonton: Edmonton Journal Group.
- Hillier, Amy, Knowles, Anne Kelly (eds.)** (2008). *Placing History: How Maps, Spatial Data and GIS are Changing Historical Scholarship*. Redlands, CA: ESRI Press, CD-ROM.
- Hypercities*. <http://hypercities.com>.
- Imagining Toronto*. <http://imaginingtoronto.com>.
- Knowles, Anne Kelly (ed.)** (2002). *Past Time, Past Place: GIS for History*. Redlands, CA: ESRI Press.
- Lippard, Lucy** (1997). 'Sweet Home and Being in Place'. *The Lure of the Local: Senses of Place in a Multicentered Society*. New York: New Press, pp. 22-38.
- London Sound Survey*. <http://www.soundsurvey.org.uk>.
- Montreal Sound Map*. <http://cessa.music.concordia.ca/soundmap/en>.
- Moretti, Franco** (2005). *Graphs Maps Trees: Abstract Models for Literary History*. New York: Verso.
- New York Sound Map*. <http://fm.hunter.cuny.edu/nysae/nysoundmap/soundseeker.html>.
- Open Street Map project*. <http://bestofosm.org/>.
- San Francisco biomapping project*. <http://biomapping.net/>.
- Sindon, Diana Stuart (ed.)** (2007). *Understanding Place: GIS and Mapping Across the Curriculum*. Redlands, CA: ESRI Press.
- Solnit, Rebecca** (2001). 'The Solitary Stroller and the City'. *Wanderlust: A History of Walking*. New York: Penguin, pp. 171-195.
- Thompson, Nato, Independent Curators International** (2008). *Experimental Geography: Radical Approaches to Landscape, Cartography, and Urbanism*. New York: Melville House.
- Turchi, Peter** (2007). *Maps of the Imagination: The Writer as Cartographer*. San Antonio, TX: Trinity University Press.
- Urban Sketchers*. <http://www.urbansketchers.com>.
- VozMob*. <http://vozmob.net/en>.
- Walk Score*. <http://www.walkscore.com>.

Text Encoding and Ontology – Enlarging an Ontology by Semi-Automatic Generated Instances

Amélie Zöllner-Weber

amelie.zoellnerweber@gmail.com
University of Bergen, Norway

In this contribution, we present an application that supports users when working with ontologies in literary studies. Thereby, semi-automatic suggestions for including information in an ontology are generated. This application is meant for users, who are familiar with annotation and markup and are interested in topic of literary studies.

When reading literature we can identify literary phenomena but we cannot prove them directly in the text. Our ability is to puzzle sentences together so that they form a meaning. But this process happens in our mind not in texts. However, these interpretations are individual and can differ from reader to reader since they are influenced by our cultural and social background. It is therefore a challenge to create a model of these interpretations to be able to have a more general and formal description, e.g. of a character.

In computer philology, one can detect several applications when modeling texts: 1) by using mark up languages like XML (meta) information can be marked in texts (e.g. Jannidis et al. 2006, Meister 2003), 2) one can model theories in literary studies that try to represent mental representations (Jannidis 2004, Schneider 2000). However, text structures and mental representations can differ from each other so that we are not able to model them in the same way.

In Zöllner-Weber 2007, mental representations have been modelled by an ontology. It tried to regard a character as a complex cognitive entity in the reader's mind. Here, the description of literary characters has been realised as an ontology. For manipulating this ontology, users have to extract information manually about characters from literary texts and add them to the ontology. This process might be time-consuming, and users who are not familiar with the structure of an ontology might need even more time to become familiar with the application. We want to solve this problem by combining text encoding and the ontology. Therefore, an annotation system has

been developed, which takes the mark up from the text and generates semi-automatically suggestions of instances be included into the ontology.

1. Methods

For the description of literary characters, an ontology that models characters by their mental representations was used (Zöllner-Weber 2006). Briefly, an ontology is a hierarchy of classes. In addition, the classes contain instances that represent individuals. Properties, which contain additional information, are attached to the individuals (Noy et al. 2001). By using this kind of structure, information is described formally. We chose an ontology because its structure corresponds to the mostly hierarchical structures of proposed theories to describe or analyse literary characters (Jannidis 2004, Lotman 1977). Several theories of literary characters are combined to create a base of a formal description using an OWL ontology (Grigoris and Harmelen 2003, Jannidis 2004, Nieragden 1995). The frame of mental representations is presented by the main classes of the ontology, e.g. inner and outer features, actions on other characters and objects. The sub classes contain characteristics of special characters (special features or groups of characters). We decided to include single pieces of information gained from literary texts into instances of the ontology. In addition, so-called instances of the classes represent individual and explicit objects of the domain of literary characters. Here, direct information about a character given in a text is assigned to an instance. Properties contain additional information, e.g. type of narrator, author, annotation information or reference to literature. Together with the information of the class hierarchy, instances and their properties, a single mental representation of a character is modelled (cf. Figure 1). In this approach, individual description, the pre-step of interpretation, is focused. The main description categories secure a general classification so that it is also possible to compare two different interpretations of one character, which might be spread over different categories of the ontology.

In order to fill the aforementioned ontology of literary characters in a more automated fashion an encoding scheme has been developed. For the annotation, we selected tags of the TEI-DTD (Text Encoding Initiative 2003, <http://www.tei-c.org/>), which were developed for marking interpretation sections in texts. Thereby, the encoding scheme had to be exploited and rearranged so that it is usable for literary studies. This means that the usage of elements was enlarged. By using this special markup, a user can directly add interpretive pieces of information about a literary character to a text. Here, the annotation scheme is based on

four main categories, *description*, *statement*, *action*, and *speech*, which classify pieces of information. All descriptions about a character that are stated by a narrator are subsumed under *description*. The category *statement* depicts commentaries of a character about another character. To mark non-verbal and verbal actions of a character, the categories *action* and *speech* should be used. In addition, a user should add e.g. information about the type of narrator, the name of a character and depending on the chosen category additional information to complete the annotation. After the process of annotation, a user sends the marked texts via a web form to a server where the annotations are evaluated by an in-house developed programming algorithm (cf. Fig 1). The pre-sorting of encoded information about a character is based on the four categories, which match the main classes of the ontology. If further encoded information is given by the markup, the algorithm tries to generate a further sub-classification. Figure 1 depicts an example of this process. After successful processing, a user is presented with a list for all processed annotations that probably form instances. Additionally, for all of these suggestions a class assignment is also given.

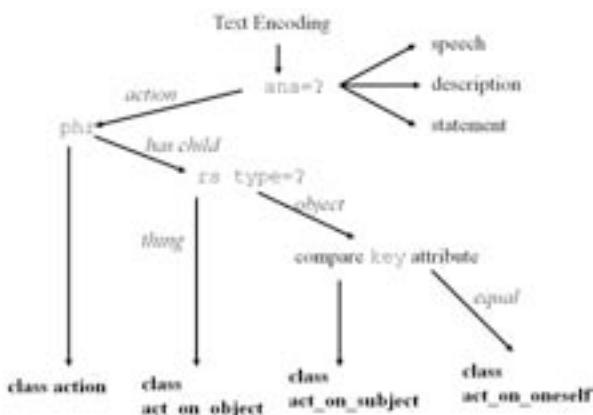


Figure 1

In addition, we present surrounding classes by showing an extracted list of classes of the ontology so that a user is able to inspect the environment of the new instance and its class. Whether a class that should include a new instance does not exist yet, a user can also add a new class. Afterwards, (s)he can include the instance in the new class.

2. Results

The application has been tested by using an extract of the novel "Melmoth the Wanderer" (1820), written by Charles Robert Maturin. We encoded the text with the mentioned TEI-DTD and afterwards, by using the programming algorithm, we obtained suggestions for new instances. In figure 2, the process of generating an instance from a text passage is shown

as an example. For the main character Melmoth 72 instances were generated and assigned to the ontology.



Figure 2

3. Conclusion

In this contribution, a system has been presented that includes information into an ontology, which is generated from markup. We tested this application by using an ontology for literary characters. In comparison to the manual manipulation of the ontology, the application comprises a semi-automatic generation of ontology instances and supports the user when assigning this information about a character to classes of the ontology. In addition, it is not only possible to add information about a single character to the ontology, but the application can simultaneously process annotations of several characters. Thereby, time and work can be saved, as the whole text can be annotated at once and will then be transferred to the ontology. There is no need to go back and forth between text and ontology as for the pure manual insertion of character information into an ontology.

Ontologies and their applications are often linked to logical reasoning. However, incorporating such techniques into the present application might be difficult, especially for untrained users, as shown elsewhere (Zöllner-Weber 2009).

References

- Grigoris, A., Harmelen, F. V. (2003).** 'Web ontology Language: OWL'. *Handbook on Ontologies*. Staab, S., Studer, R. (eds.). Berlin: Springer, pp. 67-92.
- Jannidis, F. (2004).** *Figur und Person - Beitrag zur historischen Narratologie*. Berlin: Gruyter.

Jannidis, F., Lauer, G., Rapp, A. (2006) (2006). 'Hohe Romane und blaue Bibliotheken. Zum Forschungsprogramm einer computergestützten Buch- und Narratologiegeschichte des Romans in Deutschland (1500-1900)'. *Literatur und Literaturwissenschaft auf dem Weg zu den neuen Medien*. Lucas, M. G., Loop, J., Stolz, M. (eds.). Bern.

Lotman, J. M. (1977). *The Structure of the Artistic Text*. Michigan: University of Michigan Press.

Meister, J. C. (2003). *Computing Action. A Narratological Approach*. Berlin/New York: Gruyter.

Noy, N. F., McGuinness, D. L. (2001). 'Ontology Development 101: A Guide to Creating Your First ontology'. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Schneider, R. (2000). *Grundriß zur kognitiven Theorie der Figurenrezeption am Beispiel des viktorianischen Romans*. Tübingen: Stauffenburg.

Zöllner-Weber, A. (2006). 'Formale Repräsentation und Beschreibung von literarischen Figuren'. *Jahrbuch für Computerphilologie*. 7: 187–203.

Zöllner-Weber, A. (2007). 'Noctua literaria - A System for a Formal Description of Literary Characters'. *Data Structures for Linguistic Resources and Applications*. Rehm, G., Witt, A., Lemnitzer, L. (eds.). Tübingen: Narr, pp. 113-121.

Zöllner-Weber, A. (2009). 'Ontologies and Logic Reasoning as Tools in Humanities?'. *Digital Humanities Quarterly*. 3(4).

Index of Authors

- | | |
|------------------------------|--------------|
| Aida, Mitsuru..... | 68 |
| Akama, Ryo..... | 68 |
| Al-Bahloly, Saleem..... | 181 |
| Allen, Robert..... | 235 |
| Anand, Kshitiz..... | 260 |
| Anderson, Deborah..... | 87 |
| Anderson, Sheila..... | 84 |
| Antonijevic, Smiljana..... | 44 |
| Antoniuk, Jeffery..... | 105 |
| Ashton, Andrew..... | 89 |
| Athenikos, Sofia J..... | 92 |
| Atwell, Eric..... | 103 |
| Barker, Elton..... | 94 |
| Batjargal, Biligsaikhan..... | 279 |
| Battino, Paolo..... | 52 |
| Bauer, Michael..... | 105 |
| Bauman, Syd..... | 148 |
| Bański, Piotr..... | 98 |
| Berberich, Jennifer..... | 105 |
| Beyer, Stefan..... | 126 |
| Bingenheimer, Marcus..... | 282 |
| Blaney, Jonathan..... | 100 |
| Blanke, Tobias..... | 44, 284 |
| Blumenstein, Judith..... | 126 |
| Bodel, John..... | 337 |
| Bodenhamer, David..... | 44 |
| Bohata, Kirsti..... | 108 |
| Bork, John..... | 101 |
| Borovsky, Zoe..... | 76 |
| Bos, Erik-Jan..... | 211 |
| Bouchard, Matt..... | 354 |
| Bouzarovski, Stefan..... | 94 |
| Brierley, Claire..... | 103 |
| Brown, James..... | 285 |
| Brown, Susan..... | 82, 105 |
| Buchanan, George..... | 108, 287 |
| Büchler, Marco..... | 113, 126 |
| Bunde, Janet Marie..... | 290 |
| Burns, John..... | 267 |
| Byron, Mark..... | 110 |
| Canteaut, Olivier..... | 299 |
| Caton, Paul..... | 115 |
| Cayless, Hugh..... | 372 |
| Chartier, Jean-François..... | 225 |
| Chávez, Guillermo..... | 350 |
| Clement, Tanya..... | 31 |
| Cole, Tim..... | 175 |
| Conway, Paul..... | 118, 120 |
| Coufal, Christopher..... | 123 |
| Craig, Hugh..... | 124 |
| Crochunis, Tom C..... | 130 |
| Crofts, Daniel W..... | 163 |
| Cunningham, Richard..... | 232 |
| van Dalen-Oskam, Karina..... | 37 |
| Darányi, Sándor..... | 327 |
| DeSouza-Coelho, Shawn..... | 354 |
| Declerck, Thierry..... | 327 |
| Deufert, Marcus..... | 126 |
| Dobson, Teresa M..... | 292, 354 |
| Dodd, Helen..... | 287 |
| Doorn, Peter..... | 128 |
| Duff, Wendy..... | 232 |
| Dunn, Stuart..... | 160 |
| Eberle-Sinatra, Michael... | 82, 130, 292 |
| Eckart, Thomas..... | 113 |
| Eckhardt, Kevin..... | 235 |
| Eder, Maciej..... | 132, 219 |
| Engel, Deena..... | 290 |
| Engel, Maureen..... | 387 |
| Esteva, Maria..... | 135 |
| Evans, Joanne..... | 138 |
| Farr, Erika..... | 71 |
| Finn, Edward..... | 140 |
| Fiormonte, Domenico..... | 143 |
| Fitzpatrick, Kathleen..... | 146 |
| Flanders, Julia..... | 148 |
| Flower, John..... | 150 |
| Foong, Pin Sym..... | 260 |
| Forstall, C. W..... | 294 |
| Fraistat, Neil..... | 84, 120 |
| France, Fenella G..... | 153 |
| Friedman, Scott..... | 373 |
| Fu, Liuliu..... | 156 |
| Gabriele, Sandra..... | 354 |

Galina, Isabel.....	350	Jolivet, Vincent.....	299
Gallet-Blanchard, Liliane.....	335	Jones, Steven E.....	317
Galloway, Patricia.....	120	Juola, Patrick.....	46, 123, 319, 320
Garfinkel, Susan.....	296	Juuso, Ilkka.....	194
Geßner, Annette.....	113	Kainz, Chad.....	84
Glorieux, Frédéric.....	299	Kamal, Ranaweeram.....	208
Gold, Nicolas.....	160	Kansa, Sarah Whitcher.....	181
Graham, Wayne.....	301	Kassung, Christian.....	311
Gregory, Ian.....	159	Kautonen, Heli.....	321
Guadalupi, Laura.....	143	Keating, John G.....	171
Guy, Georgina.....	160	Kelleher, Margaret.....	171
Hachimura, Kōzaburō.....	68	Kepper, Johannes.....	184
Hansen, Eric F.....	153	Keshani, Hussein.....	324
Hara, Shoichiro.....	68	Kestemont, Mike.....	37
Haswell, Eric Andrew.....	284	Khaltarkhuu, Garmaabazar.....	279
Heckscher, Jurretta Jordan.....	296	Kimura, Fuminori.....	279
Hendricks, Harold.....	61	Kirschenbaum, Matthew.....	46, 71
Henningham, Nikki.....	138	Kong, Annemarie.....	354
Henry, Charles J.....	7	Kozaburo, Hachimura.....	368
van den Heuvel, Charles.....	44, 211	Kraus, Kari.....	120
Heyer, Gerhard.....	113	Krauwer, Steven.....	84
Hirsch, Brett D.....	303	Kretzschmar, William A. Jr.....	194
Holmes, David I.....	163	Lam, David.....	354
Holmes, Martin.....	165	Lancaster, Lewis.....	185
Honkapohja, Alpo.....	304	Lancioni, Tarcisio.....	52
Hooper, Wally.....	262	Lang, Anouk.....	187
Hoover, David L.....	168, 306	Lavrentiev, Alexei.....	190
Hori, Masahiro.....	309	Lawrence, K. Faith.....	52
Hotson, Howard.....	285	Lendvai, Piroska.....	327
Howell, Sonia.....	171	Leonard, Pamela.....	150
Hug, Marius.....	311	Llewellyn, Clare.....	267
Hui, Barbara.....	174	Lobin, Henning.....	331
Huitfeldt, Claus.....	244, 313	Lucky, Shannon.....	292
Hung, Jen-jou.....	282	Lüngen, Harald.....	331
Hunter, Jane.....	175	Maeda, Akira.....	279
Imahayashi, Osamu.....	309	Malec, Scott.....	327
Inaba, Mitsuyuki.....	344, 365	Maly, Kurt.....	156
Isaksen, Leif.....	94	Marcoux, Yves.....	244, 313
Jacobson, S.L.....	294	Martin, Worthy.....	150
Jannidis, Fotis.....	31, 44	Martinet, Marie-Madeleine.....	335
Jefferies, Neil.....	285	Martínez, Alí.....	350
Jeffries, Lesley.....	178	McCarty, Willard.....	31
Jewell, Michael O.....	52	McDonald, Jarom Lyle.....	61

- Melby, Alan K..... 61
 Meredith-Lobay, Megan..... 208
 Meschini, Federico..... 206
 Meunier, Jean-Guy..... 225
 Meyer, Sebastian..... 311
 Moore, Elise..... 235
 Morgan, Helen..... 138
 Muller, A. Charles..... 68
 Mylonas, Elli..... 337
 Nagasaki, Kiyonori..... 68
 Nelson, Brent..... 339
 Nishio, Miyuki..... 309
 Nooy, Wouter De..... 76
 Nowviskie, Bethany..... 44, 192, 341
 Ogiso, Toshinobu..... 68
 Ohno, Shin..... 344, 365
 Okamoto, Takaaki..... 347
 Olson, Michael..... 71
 Opas-Hänninen, Lisa Lena..... 194
 Ore, Christian-Emil..... 196
 Ore, Espen S..... 196
 Organisciak, Peter..... 208
 Payette, Nicolas..... 225
 Pelling, Chris..... 94
 Piez, Wendell..... 202
 Porter, Dot..... 192
 Priani, Ernesto..... 350
 Przepiórkowski, Adam..... 98
 Raben, Joe..... 8
 Radzikowska, Milena..... 105
 Ramesh, Vignesh..... 260
 Redwine, Gabriela..... 71
 Rehberger, Dean..... 120
 Reside, Doug..... 353
 Rissen, Paul..... 52
 Roberts-Smith, Jennifer..... 354
 Robey, David..... 84
 Robinson, Peter..... 206
 Rockwell, Geoffrey..... 44, 82, 208
 Rodriguez, Omar..... 354
 Rodríguez Ortega, Nuria..... 199
 Roeder, Torsten..... 356
 Romanello, Matteo..... 360
 Roorda, Dirk..... 128, 211
 Ross, Claire..... 214
 Rudman, Joseph..... 217
 Ruecker, Stan..... 105, 208, 292, 354
 Rybicki, Jan..... 219, 363
 Sabin, Philip..... 46
 Sachs, Jon..... 130
 Sainio, Tapani..... 321
 Sainte-Marie, Maxime B..... 225
 Saito, Shinya..... 344, 365
 Salah, Almila Akdag..... 76
 Sanderson, Robert..... 175
 Scheirer, W. J..... 294
 Schlitz, Stephanie A..... 228
 Schlosser, Melanie..... 230
 Seiya, Tsuruta..... 368
 Seppänen, Tapio..... 194
 Serge, Heiden..... 190
 Shabout, Nada..... 181
 Shillingsburg, Peter..... 317
 Shimoda, Masahiro..... 68
 Short, Harold..... 84
 Siemens, Lynne..... 82, 232
 Siemens, Ray..... 82
 Sinclair, Stéfan..... 82, 208, 354
 Smith, Natasha..... 235, 372
 Snyder, Lisa M..... 373
 Sokół, Małgorzata..... 238
 Sookhanaphibarn, Kingkarn..... 239
 Sperberg-McQueen, C. M..... 244, 313
 Sternfeld, Joshua..... 246
 Stickney, Zephorene L..... 377
 Swanstrom, Lisa..... 249
 Tabata, Tomoji..... 68, 309
 Tarte, Ségolène M..... 251
 Tasovac, Toma..... 254
 Taylor, Karen..... 354
 Terras, Melissa..... 9, 214
 Thaisen, Jacob..... 37
 Thawonmas, Ruck..... 239, 257
 Thiruvathukal, George K..... 317
 Timney, Meagan..... 82, 303
 Toledo, Alejandro..... 257

Tomasek, Kathryn.....	377
Toth, Michael B.....	153
Trebesius, Andreas.....	126
Trettien, Whitney.....	386
Ulman, H. Lewis.....	230
Vakkari, Mikael.....	321
Van de Sompel, Herbert.....	175
Viglianti, Raffaele.....	380
Walker, Brian David.....	178
Walsh, John A.....	260, 262
Walter, Katherine.....	120
Warwick, Claire.....	214, 232
Welger-Barboza, Corinne.....	264
Welsh, Anne.....	214
Wendts, Heidi.....	337
West, Kris.....	267
Whisnant, Anne.....	235
Wiersma, Wybo.....	383
Wittern, Christian.....	271
Woong, Choi.....	368
Wu, Harris.....	156
Wynne, Martin.....	84
Xu, Weijia.....	135
Yasuoka, Koichi.....	68
Yepdieu, Adrien.....	190
York, Christopher.....	386
Yung, Terence.....	105
Zhu, Jichen.....	273
Zöllner-Weber, Amélie.....	390
Zubair, Mohammad.....	156
van Zundert, Joris.....	44
Zwicker, Heather.....	387