JADH2013 & DH-JAC2013 CONFERENCE

ABSTRACTS

19-21 September 2013

Ritsumeikan University, Kinugasa Campus,
Conference Room, Soshikan Hall, 1st Fl.

http://www.dh-jac.net/jadh2013/

Hosted by:
JADH 2013 Organizing Committee
Under the auspices of the Japanese Association for Digital Humanities

Co-hosted by:
Art Research Center, Ritsumeikan University
Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University
Center for Evolving Humanities, Graduate School of Humanities and Sociology, University of Tokyo
International Institute for Digital Humanities

Co-sponsored by:
Japan Society of Information and Knowledge
The Mathematical Linguistic Society of Japan
IPSJ SIG Computers and the Humanities
Japan Association for East Asian Text Processing (JAET)
Japan Association for English Corpus Studies (JAECS)
Japan Association for the Contemporary and Applied Philosophy
Japan Art Documentation Society (JADS)

Thursday, 19 *September* 2013

8:30-	Registration
9:00- 17:00	TEI Workshop 2013: ***James Cummings*** (Soshikan Hall 4th FL./ 405, 406)
9:15- 17:00	Historical GIS Workshop: ***Keiji Yano*** (Seishinkan Hall 2nd FL./ #526)
17:00-	**[PRE-CONFERENCE LECTURE]** Chair: A. Charles Muller ***Harold Short***, and ***Ray Siemens***, "International Communities of Practice and Global Strategies for Digital Humanities" *6*
18:30- 20:30	Welcome Drink @Calme

Friday, 20 *September* 2013

Soshikan Hall / Conference Room

8:30-	Registration
9:00-	Opening
9:30-	**[SESSION 1]** Chair: Takafumi Suzuki ***Jonathan Hope***, 'The Language of Tragedy: Tracking a Shifting Target in a Historical Corpus' *7* ***Maki Miyake***, 'Different Characteristics of Variant Readings Based on Comparison of Major Textual Similarity Measures' *9* ***Asanobu Kitamoto***, and ***Yoko Nishimura***, 'Data Criticism: A Methodology for the Quantitative Evaluation of Non-Textual Historical Sources with Case Studies on Silk Road Maps and Photographs' *10*
11:00-	Coffee Break
11:15-	**[SESSION 2]** Chair: Toru Tomabechi ***Koichi Takahashi***, 'A Consideration on Marking up the *Madhyāntavibhāgaṭīkā* by Using TEI P5: A Complicated Case of a Critical Edition Including Extensive Reconstruction' *12* ***Espen S. Ore***, 'A Nordic Tradition for Digital Scholarly Editions?' *13* ***Tomohiko Morioka***, 'Linked Open Data for Chinese Characters' *15*
12:45-	Lunch
14:00-	**[PANEL 1]** Chair: Jennifer Edmond ***Cormac Hampson, Jennifer Edmond***, and ***Susan Schreibman***, 'Semantic Uplift in the Digital Humanities' *17*

15:30-	Coffee Break
15:45-	**[PLENARY 1: DH-JAC2013]** Chair: Keiko Suzuki Speaker: ***Ellis Tinios***, ***Akihiko Takano***, ***John Resig***, and ***Ryo Akama***, "A New Stage in the Development of Cultural Resource Databases' *21*
17:15-	**POSTER SESSION** (Soshikan Hall 3rd FL./ 303, 304)
19:00- 21:00	Banquet @TAWAWA

Saturday 21 September 2013

Soshikan Hall / Conference Room

8:00-	Registration
8:30-	**[SESSION 3]** Chair: Christian Wittern ***Juan F. Belmonte***, 'Spaces Beyond the Human: An Expanded Analysis of Traversable Space in Video Games' *22* ***Jennifer Edmond***, and ***Susan Schreibman***, 'The Hub and Spoke Model of Digital Humanities Infrastructure' *23* ***J. Stephen Downie***, and ***David Bainbridge***, 'Integrating Independent Discovery and Analysis Tools for the HathiTrust Corpus: Enhancing Fair Use Digital Scholarship' *25*
9:45-	Coffee Break
10:05-	**[SESSION 4]** Chair: Espen S. Ore ***Takako Hashimoto***, and ***Yukari Shirota***, 'Framework of an Advisory Message Board for Women Victims of the East Japan Earthquake Disaster' *27* ***Christian Wittern***, 'The Daozang Jiyao Electronic Edition - Considerations for a Sustainable Scholarly Digital Resource' *29* ***Paul Arthur***, 'Online Biographical Dictionaries as Virtual Research Environments' *30*
11:20-	Break
11:30-	**[PLENARY 2]** Chair: A. Charles Muller ***James Cummings***, 'ODDly Pragmatic: Documenting Encoding Practices in Digital Humanities Projects' *31*
12:40-	JADH AGM (Lunch)
14:10-	**[SESSION 5]** Chair: Shoichiro Hara ***Yoshiaki Murao***, and ***Yoichi Seino***, 'Designing the Chronological Reference Model to Define Temporal Attributes' *32*

	Yu Fujimoto, 'Ideal Type Modelling and Analysis - A Model Driven Approach in Cultural Sciences -' *33*
15:55-	Coffee Break
15:15-	**[PANEL 2]** Chair: Susan Brown ***Geoff G. Roeder***, ***Teresa Dobson***, ***Ernesto Peña***, ***Susan Brown***, ***Elena Dergacheva***, and ***Ruth Knechtel***, 'Exploring the Future of Publishing through Software Prototyping: Three Reports on a Workflow-Management User Experience Study' *35*
16:45-	Break
16:55-	**[PLENARY 3]** Chair: Mitsuyuki Inaba ***Paul Arthur***, ***Jieh Hsiang***, and ***Masahiro Shimoda***, 'Transcending Borders through DH Networking in the Asia-Pacific' *40*
17:55- 18:20	Closing

POSTER SESSION

1. ***Yui Arakawa***, ***Ryosuke Yoshimoto***, ***Fuyuki Yoshikane***, and ***Takafumi Suzuki***
'An Investigation on Twitter Use of Researchers: User Type Classification and Text Analysis' *41*

2. ***Hajime Murai***
'Validation of Hierarchical Rhetorical Structure by Combination of Quantitative Evaluation and Traditional Rhetorical Analysis' *43*

3. ***J. Stephen Downie***, ***Tim Cole***, ***Beth Plale***, and ***John Unsworth***
'Workset Creation for Scholarly Analysis: Preliminary Research at the HathiTrust Research Center' *44*

4. ***Mamiko Mataza***, ***Akihiro Tsukamoto***, and ***Tomoki Nakaya***
'Geographical structure of medical delivery in Kyoto, Tokugawa Japan: A Historical GIS Analysis on the Distribution of Medical Practitioners' *46*

5. ***Mikio Fuse***, and ***Asanobu Kitamoto***
'Corrective-Detective Features in the Ongoing *Finnegans Wake* Genetic Research Archive Project' *48*

6. ***Miguel Escobar***
'The *Contemporary Wayang Archive*: A Digital Inquiry into the Ethics and Aesthetics of a Theatre Tradition' *50*

International Communities of Practice and Global Strategies for Digital Humanities

Harold Short (King's College London), and ***Ray Siemens*** (University of Victoria)

International associations concerned with digital humanities originated in the 1960s and 1970s, with roots in North America and Europe but with strong connections to important digital scholarship in the arts and humanities in other parts of the world, for example in Japan and Australasia. In the more recent past there has been an explosion of activity across the globe in this now rapidly growing field, and this is reflected in a number of international activities and initiatives.

Sensing the huge potential for this growth, the longer-standing digital humanities associations began discussions in 2002 that led three years later to the establishment of the Alliance of Digital Humanities Organisations (ADHO). The primary purpose of ADHO was to provide a framework for promoting and supporting the international development of digital humanities. The initial emphasis was on publications and the annual international conference, but broader aims were to do with global developments, crossing geographical, cultural and language boundaries.

Siemens will explore the methodological commons and the rapidly developing digital humanities communities and communities of practice. Drawing on his experience of the Digital Humanities Summer Institute and the remarkable range and variety of THATCamps, workshops and other training events and meetings taking place around the world, he will describe and elaborate emerging models of networked activities at local, regional and national level, and designed to be an integral part of the developing ADHO strategies.

Short will review current ADHO initiatives, including the admission of new associations and new types of association into the ADHO family, and the initiation of Special Interest Groups, the first being GO:DH (Global Outreach – Digital Humanities), which has begun in spectacular fashion. He will also describe current activity and future strategies to promote multi-cultural and multi-lingual approaches to digital scholarship in the arts and humanities around the globe.

Both talks will strongly reflect ADHO's commitment to engaging with and supporting the development of inclusive, global digital humanities.

The Language of Tragedy: Tracking a Shifting Target in a Historical Corpus

Jonathan Hope (Strathclyde University)

In a number of papers, Michael Witmore and I have used digital techniques to explore the relationship between genre and language in Shakespeare (2004, 2007, 2010, 2012). Perhaps our most significant, and surprising, finding to date has been the very strong relationship that can be identified between literary genre (which we take to be a high-level attribute of texts) and micro-linguistic features (a low-level attribute of texts).

We consider genre to be an attribute of texts which is primarily ascribed culturally: writers, critics or others can identify the genre of a text at the time of writing, or subsequent to writing. Texts can be assigned to 'new' genres unknown to their writers (as happened to Shakespeare's Late Plays at the end of the nineteenth century). Texts can be assigned to more than one genre simultaneously, and critical communities can argue fruitfully about the generic attribution of texts. These features of genre suggest that the genre of a text is in some sense not an intrinsic part of the text itself.

By contrast, the micro-linguistic features that occur in texts do so objectively, from the first writing of the text, and do not change. There is no room for argument about item frequencies unless the terms of counting are changed: if a linguistic item is reclassified, for example. Such micro-linguistic features are therefore an intrinsic part of the text.

If these views of genre and micro-linguistic feature are accepted, it is surprising to find that frequency counts of certain micro-linguistic features give very reliable predictions of genre: a high-level, apparently subjective feature of texts is closely tied to low-level, objective ones.

So far, we have considered genre within the work of one writer, Shakespeare, but we have now begun to explore the linguistic basis of genre in a multi-author corpus spread out over time. In this paper we will attempt to trace the genre of tragedy in the corpus of printed Early Modern plays (c. 400 plays from 1550-1640). We are interested in tracing linguistic evolution in 'tragedy' over the period: does the genre share a linguistic 'fingerprint' over the whole period? Is there a recognisable direction of movement through the multi-dimensional space of our analysis? In what sense is a late tragedy the 'same' as an early tragedy?

Our aim is not only to investigate the development of tragic drama in English, but to demonstrate techniques for tracking moving targets in historical corpora. One of the exciting possibilities of future digital research is the description of the development of genres over long time periods, but this raises profound ontological questions about continuity of identity; and indeed the nature of the types of identity we ascribe to groups of texts.

[Key Words: Text analytics, Early Modern English, genre theory, visualisation, data mining]

Jonathan Hope and Michael Witmore, 2004, 'The very large textual object: a prosthetic reading of Shakespeare', *Early Modern Literary Studies* 9.3 / Special Issue 12: 6.1-36 http://www.shu.ac.uk/emls/09-3/hopewhit.htm

Michael Witmore and Jonathan Hope, 2007, 'Shakespeare by the Numbers: on the Linguistic Texture of the Late Plays' in *Early Modern Tragicomedy*, Subha Mukherji and Raphael Lyne (eds), (D.S. Brewer), pp. 133-53

Jonathan Hope and Michael Witmore, 2010, 'The Hundredth Psalm to the Tune of "Green Sleeves": Digital Approaches to the Language of Genre', *Shakespeare Quarterly*, vol. 61, no. 3 (Fall 2010), pp. 357-90 http://muse.jhu.edu/journals/shakespeare_quarterly/toc/shq.61.3.html

Michael Witmore and Jonathan Hope, 2012, 'Après le déluge, More Criticism: Philology, Literary History and Ancestral Reading in the Coming Post-Transcription World', *Renaissance Drama* 40, pp. 135-50

Different Characteristics of Variant Readings Based on Comparison of Major Textual Similarity Measures

Maki Miyake (Osaka University)

In this study, we present a computational approach to textual variant analysis, focusing specially on the influential modern editions of the Greek New Testament. As the Nestle-Aland Greek New Testament is considered to be one of the most useful critical materials for current biblical studies, we set the 27th edition as the norm text. Even though this is not the latest version (the 28th edition was released last September), it is still widely recognized as the primary source. We compared the 27th edition with some modern editions such as Westcott-Hort (1881), Scrivener (1894) and Robinson-Pierpont Majority Text (2005) within the synoptic Gospels; the Gospels of Matthew, Mark and Luke are similar in structure, content, and word.

The similarity measures were conducted on a contiguous sequence of words whose sizes ranged from 4 to 10 tokens. We also did the calculation verse by verse which is equivalent to doing it line by line.

We applied some promising similarity measures to investigate the differences between the editions. The major word matching algorithms, commonly used in Information Theory, were used to calculate the difference between the words. More specifically, we made use of two string-based measures such as Levenshtein distance and Jaro-Winkler distance and Cosine similarity as a token-based measure.

Levenshtein distance is well-known as an edit-distance and it calculates a distance based on simple edit operations such as insertion, deletion, and substitution.

Jaro-Winkler distance, in turn, treats the more complex edit operations including character transpositions. Cosine similarity measures the angle between the two token sets represented as vector elements. Alongside the classical measures, we focus on a fuzzy matching technique called FCosine that was recently proposed by Wang et. al. (2011). The FCosine measure is combined with Levenshtein distance and Cosine similarity and this hybrid method is capable of approximate string matching as it takes advantage of both edit-based and token-based distance. By comparing the statistical results, we explored what aspect makes the difference of variant readings. At the same time, we discussed that each similarity measure indicates how effectively the variant's characteristics.

[Key Words: Textual Variants, Edit Distance, Cosine Similarity]

Data Criticism: A Methodology for the Quantitative Evaluation of non-textual Historical Sources with Case Studies on Silk Road Maps and Photographs

Asanobu Kitamoto (National Institute of Informatics), and
Yoko Nishimura (The Oriental Library)

This paper proposes the concept of "data criticism" and demonstrates how this concept can be applied to non-textual historical sources such as maps and photographs. Data criticism deals with the evaluation of sources in the context of historical studies, which is the same role as textual criticism. Nevertheless, the criticism of data, especially non-textual (visual) sources such as maps and photographs, has not been well studied due to the lack of methodology for quantitative evaluation. Visual sources are essential for the spatial interpretation of historical facts, but the necessity of critical methods was relatively unnoticed. We believe that this is due to the appearance of visual sources, which look like the record of facts because the creation of those sources heavily depends on objective techniques and tools such as survey and camera. We show, however, that visual sources are not facts, and their quality needs to be evaluated before using them as reliable sources, in the same way as text. Hence we use old maps and photographs as case studies to demonstrate how data criticism will bring about new perspectives on the interpretation of visual sources.

We first introduce the computational part of data criticism. We developed a suite of geo-referencing techniques for maps, including support for different map projections, one-point vs. entire registration of maps, and the preservation of linear features of maps such as streets. This process is typically performed by a built-in geometric correction in geographic information systems (GIS), such as rubber-sheeting, but we show that these techniques have rooms for improvement, and propose new algorithms using ground control points and lines, or the latitude-longitude mesh on the map. The result is integrated into web-based tools, on top of geo-browsers such as Google Maps and Google Earth, and we also developed a new interface "mappinning" (map + pinning) for performing one-point registration on-the-fly.

We then discuss the historical part of data criticism; namely, how data criticism can lead to a new interpretation of visual sources and discovery of new historical facts. First, we criticized Central Asian maps from European expeditions by Stein, Hedin and others. We quantified the distribution of errors and explained the cause of errors in the context of technical limitation of survey technology available at the time. Second, we geo-referenced the old map of Beijing, "Complete Map of Peking, Qianlong Period," created about 250 years ago. We identified mis-alignment of maps possibly due to incorrect restoration in the past, and revealed for the first time the original form of the map by connecting 203 sheets based on our proposed geometric correction method. Third, we focused on ruins located in a few regions in Tarim basin, and identified conceptual relationships between ruins that appear in expedition reports and current survey reports in different names and types. This suggests that consistent interpretation of facts is possible even when text is inconsistent across sources.

In the age of data, historical studies should incorporate and integrate larger variety of data, and we believe that 'data criticism' is a key methodology for the proper interpretation of data, and the discovery of historical facts. It is a natural extension of textual criticism that has been studied for a long time, but it focuses

on the identical central concept of historical studies; namely, the evaluation of the value of sources.
Finally, most of the maps and photographs introduced in this paper are accessible at Digital Silk Road Website http://dsr.nii.ac.jp/geography/

[Key Words: Data criticism, old maps, historical sources, quantitative evaluation, Silk Road]

A Consideration on Marking up the *Madhyāntavibhāgaṭīkā* by Using TEI P5: A Complicated Case of a Critical Edition Including Extensive Reconstruction

Koichi Takahashi (University of Tokyo)

This paper will discuss how to mark up the text called *Madhyāntavibhāgaṭīkā*, a Buddhist text composed in Sanskrit about 6th century. The modern critical edition of this work has an extremely strange appearance. The opening part of the edition reads as follows:

> *uttamajanā hi prāyaśoguruṃśraddhādevatāṃcābhyarcyakarma*supravartantaityayamapyahamuttama jananayamanuvartīMadhyāntavibhāgasūtrabhāṣyaṃcikīrsur [itijñāpanārthaṃ] tatpraṇeturvaktuścapūjā ṃkṛitvātadarthavibhāgāyaprayuktaitipratipādayannāha(*Madhyāntavibhāgaṭīkā* ed.by S. Yamaguchi, Nagoya, 1934, p.1)

The text started with the italic style, but it suddenly began to use the roman font at the end of the word "*karmasu*". However, it changed again its style into the italic just after a few lines. In this way, the italic and roman lines alternately appear in this edition which has more than 200 pages. The strange structure of this edition is a result of repairing the damaged part of the Sanskrit manuscript. One third of eachleaf is entirely lost in this manuscript, which is a single material for editing this work. However, a situation peculiar to the Buddhist literature made it possible to reconstruct the lost part of this text. Most of the Buddhist canons and exegeses had been translated into Tibetan until the 12th century. Some modern scholars attempted to restore the lost part by means of philological investigation of the Tibetan rendering. Dr. S. Yamaguchi, one of these scholars, successfully repaired the whole part of the text, and issued its critical edition in 1934. This criticale dition uses the italic style to indicate the reproduction.As a result, in his edition the italic lines repeatedly appear at random without any association with the logical structure.In other words, the edition includes the physical structure representing the damaged portion of the manuscript as well as the logical structure shown by paragraphing.

In this way, the critical edition of the *Madhyāntavibhāgaṭīkā* has extremely complicated structure. In spite of the unusual appearance, it is expected to be digitized because of its philosophical significance. XML seems to be a suitable data type for this purpose, and also the TEI P5 could make it easier to mark up the text. For example, it provides the tag sets<damage> to indicate the damaged part, and <supplied> to signify the text supplied by the transcriber or editor.These tag sets, however, seem to be prepared for more simple cases. Therefore, this paper will discuss the adequacy and efficiency of these tag sets for marking up the text with the extremely complex structure.

[Key Words: XML, TEI P5, Sanskrit, Buddhist, *Madhyāntavibhāgaṭīkā*]

A Nordic Tradition for Digital Scholarly Editions?

Espen S. Ore (University of Oslo)

Since the work on digital editions started in the different Nordic countries more than 20 years ago there seems to have developed a consensus or convergence of editorial styles for digital scholarly editions in the Nordic countries (here defined as Denmark, Finland, Norway and Sweden, for Finnish editions especially those based on material written in Swedish)[1]. In this paper I want to discuss if this is really the case and present some fairly recent digital editions from Nordic countries.

The editions still in development or finished within the last five years span texts written within a time period from the early 18th century up until recent times - and if work on medieval texts is included the start of the time span goes even further back. The texts include both published literature and unpublished private writing. One special case here is the written works of the Norwegian painter Edvard Munch.

Since 1995 the community working on Nordic Textual Scholarship has had a formal network with (now) biannual conferences and special workshops[2]. This paper looks at how this level of collaboration may have lead to a convergence of how digital editions are built - both the process of constructing a digital edition and also in how the final result is presented. Among the editions discussed in this paper are Ludvig Holberg (No/Da), Søren Kierkegaard (Da), Henrik Ibsen (No), Edvard Munch (No), Zakarias Topelius (Fi), Selma Lagerlöf (Sw) and others.

The BEE used its own system for text encoding, MECS[3], and so did the edition of the complete works of Kierkegaard[4] although here as with Wittgenstein's Nachlass the proprietary encoding was later translated into XML. Since the project start on the edition of Henrik Ibsen's Writings (HIW) TEI has been the preferred encoding system. HIW was started in 1998 and so used SGML/TEI for its first few years, but this was changed into XML/TEI with the introduction of TEI P4. All the Nordic edition projects that have started after those mentioned above have been based on XML/TEI.

A digital edition is not - or not only - classified by the text encoding that is used. For almost all the Nordic digital editions we find that they are developed as archives of digital data where the edition itself is only one of more possible views into the complete archive. The following elements are usually included, at least in the archive: images (facsimiles of manuscripts and/or printed editions), encoded transcriptions of mansuscripts and/or printed editions, new editions of the texts, usually stored as encoded text and possibly facsimiles of a new printed edition and secondary data of various kinds.

While the projects listed so far have been digitally created editions there has also been a Nordic trend towards national archives of digitally available texts whether they are published as images of text (PDF and other facsimiles) or displayed text produced from encoded text files. One of the possible Nordic twists here is that text archievs are created and maintained at a national level rather than a regional or purely institutional[5]. And for all of them the published edition is not the formal end of the project: the long time conservation of the archive is an in built part of the project. This usually means that the digital archive is stored at a national institution such as a national library or a university (universities are mostly government financed institutions in the Nordic countries) although we might find an exception with Edvard Munch's writings where the digital

1 See Dahlström, Mats and Espen S. Ore: "Elektronisches Edieren in Skandinavien" in *Geschichte der Edition in Skandinavien*, de Gruyter, to be published June 2013 - ISBN 978-3-11-031757-2

2 http://www.nnedit.org (checked April 19, 2013)

3 Huitfeldt, C.: "Multi-Dimensional Texts in a One-Dimensional Medium", Computers and the Humanities 28: 235-241, 1995, http://link.springer.com/content/pdf/10.1007%2FBF01830270 (checked April 18, 2013)

4 Preface to *Søren Kierkegaards Skrifter* (Søren Kierkegaard's Writings), electronic version http://sks.dk/red/forord-e.asp (checked April 18, 2013)

5 See for instance Arkiv for Dansk Litteratur (Da): http://adl.dk/adl_pub/omadl/cv/OmAdl.xsql?nnoc=adl_pub, bokselskap.no (No): http://www.bokselskap.no/ and The Swedish Literature bank (Sw/Fi): http://litteraturbanken.se/#!om/inenglish (all checked on April 18, 2013)

archive is stored at the Munch Museum.

As far as there has been a convergence in digital editions in the Nordic countries it is that a) certain large editorial projects which may have national or other funding develop complete editions of selected authors and which build up complex archives of text and images while b) at a national level there are institutions which provide long time archives for digital editions whether they are text editions or facsimile editions. How far this represents a particular Nordic school of digital editions depends on what is done elsewhere. This paper will compare the situation in the Nordic countries with some other Western European countries and North America.

[Key Words: Digital editions, textual scholarship, text encoding, text archives]

Linked Open Data for Chinese Characters

Tomohiko Morioka (Kyoto University)

This report describes RDF mapping of the CHISE character ontology[1], especially it focuses representation and usage of Chinese Characters (Kanji, Hanzi, etc.). Character is a basis of various kinds of data.
Chinese Character is a logogram, each character is basically indicates a morpheme. In the point of view, a Chinese Character can be a hub of linguistic and semantic information about various words and texts written by Chinese characters.

The CHISE character ontology is a large scale character ontology which includes 238 thousand character-objects including Unicode and non-Unicode characters and their glyphs, etc. It was developed for CHISE (Character Information Service Environment)[2] which is a character processing system not depended on character codes. The framework of CHISE is based on database processing. It uses CONCORD as the database engine. CONCORD is a prototyping style object oriented database system based on directed graph, like RDF. We developed a Web service to display and edit objects of CONCORD, called "EsT". EsT was originally designed a HTML based Web application, however we added RDF mapping and implemented RDF/XML output feature.
Thus the CHISE character ontology can be used as a Linked Open Data.

As demonstrations of data-linking between character ontology and other information resources, I made two examples: (1) links between Oracle Bone characters and rubbing/photos of their sources (e.g. [fig.1][3]) (2) links between characters and "Bibliography of Oriental Studies"[4] (e.g. [fig.2]). In a view of character ontology, these links work as sources/examples of characters to describes and indicates characters clearly. In addition, they also work as entrance of other information resources.

CHISE provides search service for Chinese characters, named "CHISE IDS Find"[5]. If a user specifies one or more components of Chinese character into the "Character components"window and run the search, characters include every specified components are displayed. It is a very useful service to search Chinese characters, especially for characters which are difficult to input by ordinary input-methods designed for daily-used characters. Result page of the CHISE IDS Find has links for character objects using EsT (the first column of each line is a link for EsT page to display detailed information of character object). Thus users can trace the links for other information resources via EsT. It is also useful for search robots and third-party's Web services based on Web API. In the point of view, RDF/XML feature of EsT can provide deeper cooperation for third-party's applications. In an ordinary HTML page of EsT, there is a RDF button (right side of the top line). To click the RDF button, EsT outputs RDF/XML format instead of HTML. [fig.3]

As a future plan, I'm planning to make links between Chinese characters and morphemes of classical Chinese. It will be a example to integrate character processing layer and linguistic processing layers. This report focuses the concepts of the CHISE character ontology and basic mechanism of its RDF mapping briefly and describes possibility of it as a information hub for other kinds of information resources.

1 Tomohiko Morioka, "CHISE: Character Processing based on Character Ontology", Large-scale Knowledge Resources (LKR2008), pp.148-162, LNAI 4938.

2 http://www.chise.org/

3 Catalogue of the Oracle Bones in the Kyoto University Research Institute for Humanistic Studies, http://chise.zinbun.kyoto-u.ac.jp/koukotsu/

4 http://ruimoku.zinbun.kyoto-u.ac.jp/ruimoku7/

5 http://www.chise.org/ids-find

[Key Words: Chinese Character, RDF, Linked Open Data, character ontology, character representation]

[fig.1] http://www.chise.org/chisewiki/view.cgi?character=rep.zinbun-oracle:204 This source is a 3.

[fig.2] http://www.chise.org/chisewiki/view.cgi?character=ruimoku-v6:0xE000

[fig.3] http://www.chise.org/est/rdf.cgi?character=%E8%AA%AC

Semantic Uplift in the Digital Humanities

The rise in digital humanities research has led to a huge increase in the digitisation of cultural heritage artefacts. However, the resultant data often resides in independent silos that do not share their rich information or connect to other relevant collections. This lack of connectivity reduces the impact of the digitisation process and makes it more challenging to find potential synergies between dispersed artefacts. A major obstacle in providing tangible links between digital collections is the proliferation of unstructured text and insufficiently rich metadata. Trinity College Dublin is directly addressing this challenge on a number of fronts, as exemplified by its leading role in three European Commission funded projects; CULTURA[1], CENDARI[2] and DigCurV[3]. This panel session will discuss the latest research on semantic uplift (best practice education of curators, entity extraction, ontology mapping etc.) that is being performed in Trinity College Dublin, and explore its relevance to the wider digital humanities community.

In the strictest sense, semantic uplift is the process of enriching digital content with metadata and exposing it to the greater web of data. This can involve a variety of processes such as extracting entities from unstructured text, mapping the entities to domain specific ontologies, and publishing this enriched information as linked data. But the processes required to ensure digital content can be identified and interrogated by users is not always a purely technical one. Ensuring that common standards are agreed and followed, that technical developments don't exacerbate disparities in the analogue record, and that institution staff gain access to the necessary training to understand these new methodologies, and the rich potential they offer, are obstacles on this path. Trinity College Dublin is helping institutions tackle this problem through its participation in the DigCurV project. The DigCurV project is addressing the availability of vocational training for digital curators, with a focus on the new skills and competences (such as the technologies that can enrich the semantics and exposure of their data), which are essential for the long-term management of digital collections. A key outcome of the DigCurV project is the establishment of a curriculum framework from which training programmes can be developed in future. Such initiatives are central to ensuring the technical literacy of digital curators, and ultimately the improvement of how cultural collections are stored, connected and maintained.

As a wealth of cultural archives have already undergone the digitisation process, techniques are necessary to enrich these existing collections through semantic uplift. Two Trinity College led projects in the Digital Humanities are addressing this challenge directly. CULTURA is using sophisticated natural language processing and normalisation techniques as a means of adding additional structure to digital archives. Importantly, this approach is being tested on resources such as the 1641 Depositions[4], which contain archaic language and have huge inconsistencies in spelling and structure. The lessons learnt in applying semantic uplift to such challenging artefacts are highly relevant to current research directions in digital humanities.

By enriching content with more structure it significantly eases the difficulties of connecting data to other relevant external resources. This linking process is a key focus of CENDARI, a broad-based infrastructure project, where ontologies are being used to facilitate mappings, and multi-lingual resources from medieval and modern European history are ultimately being matched to concepts on the semantic web. The recent proliferation of published Linked Open Data by digital humanities organisations (such as Europeana[5]) highlights the importance of improving such processes, and the significance of semantic uplift to digital humanities practitioners.

1 http://www.cultura-strep.eu/
2 http://www.cendari.eu/
3 http://www.digcur-education.org/
4 http://1641.tcd.ie/
5 http://pro.europeana.eu/linked-open-data

‘Enriching Metadata in Digital Archives to Improve Analysis, Exploration and Connectivity’

Presenter: ***Cormac Hampson*** (Trinity College Dublin)

Digital archives typically contain metadata that describe each resource in varying levels of detail. Unfortunately, the richness of this metadata is often not sufficient to provide more than a basic classification of each document, which limits the analysis, exploration and connectivity possibilities for an archive. Furthermore, in large collections that have thousands and millions of resources, it is not practical to manually mark up each individual resource. Thus, there is a major challenge to improve the granularity of the metadata describing cultural documents, as well as ensuring that this process can scale to large digital archives. The Knowledge and Data Engineering group at Trinity College Dublin in Ireland, are leading the EU funded CULTURA project[6] which is tackling this issue through semantic uplift.

In terms of CULTURA, semantic uplift centres on natural language processing techniques. Specifically, IBM’s LanguageWare is being used to perform entity and relationship extraction over unstructured text collections. The output of this processing is an entity graph, which contains finely grained metadata describing the main entities (people, locations, dates etc.) in each resource. A key CULTURA technology in terms of semantic uplift is the PreMapper tool, which enables an expert to view and edit the exported entity graph, and perform such tasks as entity merging, entity disambiguation and entity creation. By supporting such human intervention, the PreMapper tool helps improve the accuracy of the overall entity extraction process.

Importantly, the semantic uplift approach in CULTURA is being tested on resources such as the 1641 Depositions[7], which contain archaic language and have huge inconsistencies in spelling and structure. For such challenging archives, the entity extraction process in CULTURA is helped considerably by a preceding text normalisation phase. The primary purpose of this normalisation is to produce documents without historical variation. The elimination of such variations results in documents that are more easily processed than the originals, thus improving accuracy in the semantic uplift process.

The rich metadata produced through semantic uplift is used by CULTURA to power a number of important services that support exploration, analysis and connectivity. These services include social network visualisations, recommender systems and entity oriented search; and it is the fine granularity of the metadata that makes these services so useful to users of the CULTURA portal, whether professional researchers or members of the public. Importantly, if an archive has entities extracted that are referenced on the semantic web, it is possible to map between these entities so that external information on that entity can be incorporated into the CULTURA environment. For example, the 1641 depositions contain many place names which have been mapped to the online geonames[8] database. This enables depositions from the 1641 collection to be mapped to a specific geographical area and visualised on in interactive display. Furthermore, by creating links between CULTURA collections and the wider web, it helps to increase traffic to archives and provide a richer landscape for digital humanities research. To summarise, this presentation will highlight the metadata challenges that exist in the digital humanities domain, and discuss how CULTURA is addressing these needs though a range of innovative technologies.

[Key Words: Digital Humanities, Entity Extraction, Semantic Uplift, Normalisation, CULTURA]

6 http://www.cultura-strep.eu/

7 http://1641.tcd.ie/

8 http://www.geonames.org/

‘The Taste of ‘Data Soup’ and the Creation of a Pipeline for Historical Research’

Presenter: ***Jennifer Edmond*** (Trinity College Dublin)

It is possible that the invisible algorithm behind the ‘Google Box’ has spoiled us all for appreciating the complexity of the world of data that surrounds us. As such, many historical researchers believe that the greatest barrier to reaching the fabled country of an optimised platform for historical research is merely to make digital copies of analogue research objects and make them all somehow, anyhow, available in the internet. Unfortunately, the widespread siloisation of digital humanities projects, the disappointing statistics regarding their reuse, and the limitations of projects like Europeana to support depth as well as breadth in object description, all point toward a much harsher reality. Within this landscape, the CENDARI project[9] has emerged with a set of defining principles aimed at advancing both the community’s awareness of what they need, and presenting a model for how these needs can be met.

The goals of the CENDARI project therefore cross a number of traditional boundaries – between content holders and content users, between technology and humanities, but also between the digital and the analogue worlds of scholarship. As such, CENDARI has committed to:

1. Creating a robust ‘enquiry environment’ capable of supporting far more than search and browse functionality, and doing so at a point in the research process where it does not specifically determine how that process must unfold
2. Resisting existing inequities in digital provision among national and regional systems. To merely enhance and add functionality to already well-established digital collections like those of the Bibliotheque National de France and the Imperial War Museum would not only be a failure, but a potential betrayal of the historical record, making well-known collections even easier to work with, while other collections and perspectives are left ever less accessible
3. To integrate at a deep level within the project the analogue processes known and trusted by scholars and archivists, in the knowledge that the digital environment can only supplement, rather than supplant, them.

Each of these commitments carries distinct challenges, but among the most critical is the semantic challenge of creating meaning out of ‘data soup’. This term was coined to describe the massively heterogeneous nature of the information we are expecting the CENDARI technical infrastructure to hold. In part, this is a matter of divergent technical and content holder standards e.g. data from an archival EAD or ISAD system will differ in its emphasis and fields from a library’s METS or MARC2. However, even within born digital projects, for example Europeana and e-Codices, standards have been developed with a very different set of goals and norms. Furthermore, the original objects will be highly divergent – from posters to army records to manuscripts – and many objects we might want to integrate may be in different languages and pre-digested in some way, such as a scholar’s database of a particular corpus of letters. How CENDARI will approach making this into a comprehensible, navigable and confidence-inspiring environment for the production of scholarship will be the core of this presentation.

[Key Words: Digital humanities, research infrastructure, digital archives, medieval history, modern history]

9 www.cendari.eu

‘Next Generation Training for Digital Curators’

Presenter: ***Susan Schreibman*** (Trinity College Dublin)

The DigCurV project[10] commenced in January 2011 with the remit to develop a Curriculum Framework that would meet the training needs of digital curators in the library, archives, museums and cultural activities sector.

New jobs are emerging for digital curators across Europe and Internationally, but evidence shows employers face recruitment difficulties due to skills shortages and increased demand from staff for vocational training in the field. DigCurV established a multilateral network under the Leonardo da Vinci programme to support and extend vocational training. The project addressed the availability of training for staff who wished to develop new skills needed for long-term management of digital collections, which were established in cultural institutions as a result of the i2010 strategy. These skills include management and quality assurance; professional conduct; personal qualities; and knowledge and intellectual abilities.

Trinity College Dublin was one of six organisations from five countries, each with a strong track record in the field of digital libraries and digital preservation, who were part of the DigCurV project. During the 30 months of the project, DigCurV identified, analysed and profiled existing training opportunities and methodologies, surveyed training needs, and identified the key skills and competences required of digital curators. This survey and analysis took place over the course of the first 18 months of the project, during which all partners hosted focus groups within their countries to determine the needs of the sector. It was important that the Curriculum Framework would not prescribe a curriculum or training programme, per se, but enable trainers and trainees to develop a curriculum that provided the required training, or recruitment evaluation tool. In particular, because new technologies are applied to digital collections on a regular basis (such as those that enrich semantics and create links to other content) it meant that the Curriculum Framework must enable training to evolve over time.

In addition to developing the Curriculum Framework, DigCurV also developed the CURATE! Game. The board-game was developed by TCD and MDR Partners, with the format and structure of games such as “Monopoly” or “The Game of Life”. Players work their way around the board, passing through various ‘stages’ of the digital curation project process (‘Develop - Educate - Manage’’). Through focus groups and investigations into the uses of games for educational and training purposes, DigCurV developed a prototype game, which was presented at various conferences such as DISH2011. The game has since been played across the EU, as well as the USA and Australia, and translated into Dutch, Hebrew and Lithuanian.

By the end of the project in June 2013, DigCurV had developed a curriculum framework, available both in print and online, from which training programmes can be developed. Furthermore, the DigCurV network now consists of over 450 members working in Digital Curation and Preservation related areas across the EU and Internationally. This presentation will highlight the need for digital curators to have a wide range of skills to perform their work effectively, and discuss how DigCurV’s curriculum framework can be used to significantly improve the training process.

[Key Words: Digital Humanities, Curriculum, Training, Curators]

10 http://www.digcur-education.org/

DH-JAC2013 SPECIAL SESSION

A New Stage in the Development of Cultural Resource Databases

Ellis Tinios (University of Leeds), ***Akihiko Takano*** (National Institute of Informatics),
John Resig (Dean of Computer Science, the Khan Academy), and
Ryo Akama (Ritsumeikan University)

Hitherto institutions such as libraries, museums and universities developed cultural resource databases to catalogue and manage materials in their possession. A logical progression from this start has been to link these institutional databases to form union catalogues. At the same time, specialists in certain subject areas have created very detailed but narrowly focused databases to support their own work. Researchers in the humanities have led the way in the creation of both the personal and institutional databases.

Recently new capabilities have been developed that possess significant implications for future database development and use: associative search systems, image matching systems and the like. These technologies have moved beyond the experimental stage and are now ready for practical application. They promise a higher level of functionality than was previously possible in databases.

Researchers in the humanities must now reassess their approach to archiving and data management in order to make the most of the new functionalities made possible by these developments.

This panel brings together two scholars in Japanese culture studies who are heavy users of databases and two leading developers of the next generation of databases to explore the future prospects.

Spaces Beyond the Human: An Expanded Analysis of Traversable Space in Video Games

Juan F. Belmonte (University of Murcia)

This paper studies virtual spaces in video games that resist traditional ideological analyses and proposes a theoretical framework for their study. First, it briefly reviews existing research on space in games (Frasca 2003, Jenkins 2004, Juul 2002, Nitsche 2008) in order to show that most available research focuses on the navigation of very specific types of space by humanoid avatars with easily recognizable gender and racial marks. The popularity in studies of urban space of the cities of the *Grand Theft Auto* series and their protagonists is a prime example of this.

However, what can be said about space in terms of ideology and power if we remove the human avatars/navigators from games? What happens to analyses of virtual spaces if their connection with our own physical and social spaces becomes greyed out or if the gameplay does not allow or encourage processes of consumption, destruction of, or domination over other agents in the game? This paper answers these questions by looking at three different examples that deal with space as well as its navigators differently: *flOw*, *Flower*, and *Journey*. *flOw* offers a great example for the study of the connection between space, avatars and ideology in that, at first sight, both the setting and its protagonists are hardly relatable to associations to the human. Despite its non-human actants, *flOw* does readily reproduce through its game mechanics some of the most basic ideas of Capitalism: depredatory desire and consumption as the main form of survival and advancement. And in these respect it is the most conventional of the three. *Flower* also provides players with non-human avatars (petals carried by air currents) but its spaces are partially connected to human, and sometimes semi-urban, societies (e.g. windmill fields). It is, however, the relation to other petals and flowers what remains outside of normative forms of relationship between human agents. This is because *Flower* promotes a form of liquid unity based on the collective. Finally, Journey, while having a humanoid avatar, also proposes a relation with the world that seems to circle around Eve Sedgwick's notion of 'beside' (2004). This originates from the genderless nature of the avatar as well as its material relation with the different elements of the game world such as the player-controlled avatar, the sand, the snow or the wind. The ultimate questions this paper seeks to answer are: How can virtual space be analyzed in games where the human, or at least some of the main features that have come to define the human, have been removed or intensely modified? Can we talk about the human in relation to seemingly non-human avatars and/or spaces?

[Key Words: Virtual Space, Video Games, Identity, Ideology, Avatar]

References

Frasca, G. (2003). Sim Sin City: some thoughts about Grand Theft Auto 3. Game Studies 3 (4). Retrieved from: http://www.gamestudies.org/0302/frasca/

Jenkins, H. (2004). Game Design as Narrative Architecture. In Pat Harrington y Noah Wardrup-Fruin (Eds.), *First Person: New Media as Story, Performance, and Game*. Cambridge, Massachusetts: The MIT Press.

Juul, J (2002). The Open and the Closed: Game of emergence and games of progression". In Frans Mäyrä (Ed.), *Computer Games and Digital Cultures Conference Proceedings*, Tampere: Tampere University Press, 323-329.

Nitsche, M. (2008). *Video Games Spaces*. Cambridge, Massachusetts: The MIT Press.

Sedgwick, E. K. (2003). T*ouching Feeling: Affect, Pedagogy, Performativity*. Durham & London: Duke University Press.

The Hub and Spoke Model of Digital Humanities Infrastructure

Jennifer Edmond, and ***Susan Schreibman*** (Trinity College Dublin)

The most common image deployed in discussions of cyber-infrastructure is that of the road: concrete (literally!), linear, public and established. But the applications of this metaphor usually act as a shorthand for a desired essence of availability and accessibility, ignoring both the lessons the original image can teach us and the actual properties of a modern digital research infrastructure, which are so different from those of road transport. In particular, the European Digital Research Infrastructure for the Arts and Humanities (DARIAH) and its constituent national partners and formational projects, can be far more easily described - and productively understood - through the application of a hub-and-spoke model, such as is the basis for many modern airport/airline relationships. Within this new metaphor come a number of intriguing insights and issues for digital infrastructure, including the ideas of directionality, dimensioning, up- or downgauging, ancillary services and hub bypass.

Directionality in the airline context refers to the efficiency of connections between incoming traffic and ongoing connecting traffic. A research infrastructure is also interested in traffic, but in this case it is the flow of ideas (ie knowledge exchange), rather than passengers. In particular with a business model such as DARIAH has, which is supported by smaller national partner cash contributions and much larger in kind contributions, value will only be gained by partners if they feel that there is an effective and efficient flow of information into the Hub and back out to them from other areas. This is a great challenge in knowledge enterprises in the humanities generally, where times to publication are traditionally long, and publication avenues for non-traditional forms of output are not always effective or valued. How can a research infrastructure effectively meet the challenges of directionality among its partner spokes, without infringing on institutional or individual IPR, but also without taking a merely reactive approach which does not necessarily keep the flow of knowledge moving toward its final, transformative destination?

In the airline/airport model, proper dimensioning represents a productive equilibrium between the overall incoming and outgoing capacity of people and baggage. Dimensioning ideas applied to a digital infrastructure works similarly, in analysing the amount of bandwidth needed to support users and the objects they access, both at the inward (preservation) and outward (access) service points. The efficiency of a digital infrastrastructure can therefore be measured according to how it achieves a balance in its services and approaches to services (client or server side), processing time and power, user expectation for when and how those services will be delivered, who will pay for these services, and how to develop a pricing model for them.

Up- or downgauging occurs mainly in the context of freight, where a larger body of items is transitioned at the hub onto a larger or smaller aircraft, more suited to the total volume of cargo to be transported between the Hub and a given destination. Technical developments too sometimes come in packages that are too large for other projects to use in their entirety. How can a Hub facilitate the further uptake of parts of a larger project or set of workflows elsewhere, both in terms of modularisation services but also the legal agreements needed for reuse of outputs from collaborative initiatives which may now have been disbanded?

The area of Ancillary services is another opportunity that airline hubs are able to take particular advantage of due to their high traffic volume, and status as a place where people will spend more time in the airport itself (rather than passing through on the way from landside to airside and vice versa). If knowledge is passing through a research infrastructure, therefore, there may be opportunities to standardise it, enhance it or otherwise sell it a bottle of perfume before it moves on to the next user. What might these opportunities be, and how could DARIAH and its partners take advantage of them?

Hub bypass is another issue which is worth considering in a digital humanities research infrastructure context. Sometimes, an airline finds it has enough traffic between two points that use of the hub is economically disincentivised. In these cases, normally a direct service between those points may be started, which increases effective transfer of passengers without the use of the Hub. In the digital humanities context too, there may also be times when specific projects or knowledge assets are so complementary that a strong bilateral may be a better model than a hub-and-spoke. But this should be the kind of opportunity that a digital infrastructure can foster and support, rather than be viewed as a (commercial) threat.

This model will form the basis of a theory of the construction of the DARIAH infrastructure through its central hubs and spokes, including the DARIAH Coordination Office, national members at the governmental level, institutional participants, and individual EU funded research projects closely aligned to support DARIAH's mission, including the specific example of the Collaborative European Digital Archival Research Infrastructure, or CENDARI, project.

[Key Words: Digital Infrastructure, DARIAH, Cyberinfrastrucure]

Integrating Independent Discovery and Analysis Tools for the HathiTrust Corpus: Enhancing Fair Use Digital Scholarship

J. Stephen Downie (University of Illinois at Urbana-Champaign), and
David Bainbridge (University of Waikato)

The HathiTrust Research Center (HTRC) is the research arm of the HathiTrust (HT). The HT corpus contains over 10 million volumes that comprise more than 3 billion pages drawn from some of the world's most important libraries. Founded in 2011, the HTRC is a unique collaboration that is co-located within two major institutions: University of Illinois and Indiana University. In this poster/demonstration, we introduce a new tool designed to enhance the ability of scholars to perform analyses across both the open and copyright-restricted data resources found in the HT corpus.

Approximately 69%, or nearly 7 million volumes, of the HT corpus is under copyright restrictions. To allow for fair use analytic access to restricted works, the HTRC has developed a "non-consumptive research" model. This model provides secure fair use analytic access to large corpora of copyrighted materials that would be otherwise unavailable to scholars. In our model, analytic algorithms are run against a secured version of the corpus on behalf of scholars. Submitted algorithms are vetted prior to execution for security issues, and following execution only the analytic results are returned to the submitting researchers: scholars cannot reassemble the individual pages to reconstruct copies of the original restricted works.

While the non-consumptive model is a significant step forward for allowing fair use analytic access to hitherto unavailable works, its algorithm-to-data constraints do inhibit the rapid and thorough prototyping of novel algorithms. To mitigate this shortcoming, we have combined the download and discovery functionality of the Greenstone digital library system [1] with the workflow functionality of the Meandre analytic environment [2] to operate in conjunction with the HT's bibliographic and data APIs. Using our new tool, scholars can create their own unique subsets from the HT corpus. For each work selected—both open and restricted—a bibliographic record is downloaded to their local environment. When a work is in the public domain (at the scholar's location), our tool can automatically download associated page images and text. We refer to this assembled content as the "prototyping workset." This approach helps scholars prototype new algorithms by affording them local access to as much full-text content as allowed by law, along with whatever other data to which they might have access rights unique themselves (i.e., novel annotations, locally-owned content, etc.). Once a scholar is satisfied with the behavior of their prototype, it can be submitted to HTRC to be run against the larger set of both open and restricted works.

Figure 1 illustrates a worked example of a scholar studying language evolution in New Zealand. Bibliographic records for 7500 HT works associated with NZ have been located and downloaded using Greenstone's "download-from" plugin-architecture [3]. Figure 1 shows some of the 193 documents returned after a date limitation operation (1900–1949). Note entry number six (marked with the "public domain logo"): this document will be used for local prototype analyses. Above the returned documents, a drop-down list of available Meandre workflows is provided, from which a scholar has selected a time-based tag-cloud. After iteratively modifying this workflow, and then testing it on the prototyping workset, the finalized algorithm

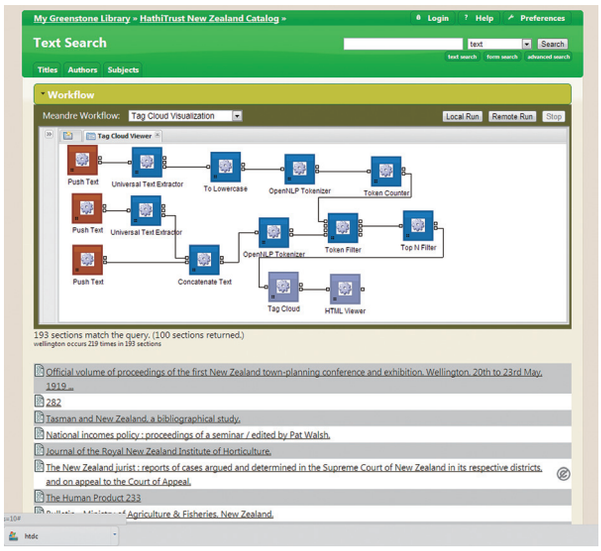


Figure 1: Screenshot showing the integration of a HT data subset and a Meandre analytic workflow within Greenstone installed on a scholar's local machine.

is then submitted to the HTRC system for vetting and subsequent execution. We believe that this "local prototype/remote execution" approach will enhance efficiency for both scholars and HTRC personnel alike as together they strive to realize the potential of fair use digital scholarship conducted against the invaluable resource that is the HT corpus.

[Key Words: Content analysis, Text analytics, Non-consumptive research, Prototyping, Workset modeling]

References

[1] Witten, I. H., Bainbridge, D., Nichols, D. M. (2010). *How to build a digital library*. Burlington, MA: Morgan Kaufmann.

[2] Llorà, X., Ács, B., Auvil, L., Capitanu, B., Welge, M., and Goldberg, D. (2008). Meandre: Semantic-driven data-intensive flows in the clouds. In *Proceedings of the Fourth International Conference on eScience*, pp. 238-245.

[3] Bainbridge, D., Thompson, J. and Witten, I. H. (2003). Assembling and enriching digital library collections. In *Proceedings of the Joint Conference on Digital Libraries*, pp. 323–334.

Framework of an Advisory Message Board for Women Victims of the East Japan Earthquake Disaster

Takako Hashimoto (Chiba University of Commerce), and
Yukari Shirota (Gakushuin University)

After the East Japan Great Earthquake, women victims have been suffering from many problems and worries: they are caring of elders, raising children, finding jobs, needs for female-specific items and so on. Administrative authorities want to recognize these women sufferers' specific problems and give them appropriate and timely advisory information. However, in general, it is difficult to grasp their needs timely because their environments and conditions change from moment to moment, and take a lot of labor for interviews or questionnaire investigations. Therefore, if we could grasp victims' requirements from social media related to the East Japan Great Earthquake by a temporal data analysis, it would be quite useful. To properly an alyze and detect thechanging needs from their messages, we set the goal of our research is to construct anadvisory message board system on web that can classify their requirements. If we can detect the hidden topics of the requirement, we will be able to give victims appropriate advices. Even if we cannot help victims, at least, we could notice their requirements. Then we will make contact with a suitable non-profit organization, and the advisory message will be also sent to the client.

To achieve our research goal, this paper aimed to develop the data mining engine which analyzes the victims' requirements. The data mining engine analyzes and detects the current topics from the sufferers messages posted on social media societies. With data analyzing results, the system can detect victims' needs transitions.

We have already acquired the social media data from the non-profit organizations[1] and developed the prototype engine for detecting their needs structures from the data.

In our previous method, victims' needs were shown as topics by adoptingthe graph-based topicextraction method[2] using the modularity measure[3]. Our previous method could show the timeseries topic structure by network graphs. But sometimes we could not extract appropriateneedsfrom graphs, because the boundaries of topics were ambiguous and subgraphs divided by the modularity could not express topics clearly. To address the problem,in this paper, we use an approach to extract hidden topics over time from socialmedia messages using the latent semantic analysis (LSA) technique[4]. In this method, first, we crawl messages in social media. Next, we construct a snapshot document-termmatrix at each time stamp. Hidden topics are extracted by using LSA for each snapshotmatrix. Then, we investigate hidden topic transitions over time. By our method, we can successfully obtain the hidden topics..There we considered afflicted people'sneeds as hidden topics extracted by the method.

In the paper, we present the framework of the advisory message board system and describe our method to extract hidden topics over time from social media messages using the latent semantic analysis (LSA) technique.

[Key Words: East Japan Great Earthquake, women disaster victim, needs transition, data mining, advisory message board]

References

[1] SAVE IWATE, http://sviwate.wordpress.com/in-english/

[2] Hashimoto, T., Kuboyama, T., Chakraborty, B., Shirota, Y., Discovering Topic Transition about the East Japan Great Earthquake in Dynamic Social Media, *Proc. Of GHTC 2012*, 259–264, 2012.

[3] Newman, M. E. J., Modularity and community structure in networks, *National Academy of Science USA* 103(23), 8577–8696, 2006.

[4] Landauer, T. K., Dumais, S. T., A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge, *Psychological Review*, 104(2), 211Ð240,1997.

[5]Hashimoto, T. and Chakraborty, B., Temporal Awareness of Needs after East Japan Great Earthquake using Latent Semantic Analysis, *Proc. of EJC 2013*, 214-226, 2013.

The Daozang Jiyao Electronic Edition – Considerations for a Sustainable Scholarly Digital Resource

Christian Wittern (Kyoto University)

The [Daozang Jiyao Electronic Edition] is part of the result of an ongoing research project of a worldwide network of more than 60 scholars on Daoism during the Qing Dynasty with a special focus on the biggest collection compiled during that period, the Daozang jiyao 道藏輯要 , 'Essentials of the Daozang'.

It has been developed in close interaction with the researchers using the collection and strives to model the text in a way that is both faithful to the complicated textual tradition and useful for modern scholarly investigation.

In this presentation, I will first give an overview of the methodological approaches to the digitization of the collection and how they have been put into practice. I will then introduce the textual model developed for this collection and discuss what potential it has beyond this research project. The next part of the presentation will introduce the website, which is currently in the beta-phase of development, to the audience as one part of the publication of the digital edition.

One of the important questions every digitalization project has to answer is the question about the long-term perspective of the project after the initial period of funding has ended. As one possible vision for digital publication, I will introduce the idea of a repository (or even better, a network of repositories) for digital editions, that integrates well with other texts published in the same way and can thus form a source for personal digital repositories maintained by scholars themselves and containing exactly the texts that are needed.
This will overcome serious problems with the current mainstream form of digital publication, namely a website as the sole venue of publication1[1]. This topic has been discussed for some time and valid suggestions and a discussion of the requirements can be found in [1], [2], [3] and [4], which this presentation will take up and expand, namely by adding the requirement that the text will not only be made available to the scholarly community, but that it also be able to annotate it in a way that can be owned by the scholar adding the annotation and still be shared with interested colleagues. A proposal for a way to implement this will also be introduced and discussed.

[Key Words: Text encoding, text representation, scholarly collaboration]

References

[1] Robinson, P. (2009). “Towards a Scholarly Editing System for the Next Decades”. In: Sanskrit Computational Linguistics: First and Second International Symposia Rocquencourt, France, October 29-31, 2007 Providence, RI, USA, May 15-17, 2008 Revised Selected Papers. Springer London, Limited, pp. 346–357.
[2] Schmidt, Desmond and Robert Colomb (June 2009). “A data structure for representing multi-version texts online”. In: International Journal of Human-Computer Studies 67.6, pp. 497–514.
[3] Shillingsburg, Peter (2010). “How Literary Works Exist: Implied, Represented, and Interpreted”. In: Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions. Ed. by Willard Mc- Carty. Cambridge: Open Book Publishers.
[4] Siemens, Ray et al. (2009). “It May Change My Understanding of the Field: Understanding Reading Tools for Scholars and Professional Readers”. In: DHQ: Digital Humanities Quarterly 3.4.
[5] Wittern, Christian (2013) "Beyond TEI: Returning the Text to the Reader". In: Journal of the Text Encoding Initiative [[http://jtei.revues.org/691]], 2013, 4.

[Daozang Jiyao Electronic Edition] http://www.daozangjiyao.org/dzjy/texts/dzjy

1 For a further discussion of this problem and a model for overcoming it see [5]

Online Biographical Dictionaries as Virtual Research Environments

Paul Arthur (University of Western Sydney)

Online biographical dictionaries and related digital resources are supporting new research methods and enabling new findings in fields of biography, prosopography, genealogy and family history. Many long-running national biographical projects, such as the Australian Dictionary of Biography and the Oxford Dictionary of National Biography, have been migrated online in the past decade. Motivations for moving from print to digital have included enabling greater user access, more affordable publication formats, added flexibility in the way information is presented, the capability to correct and update efficiently, and allowing users to locate information more quickly with the promise of greater accuracy. The transition has also been driven by policies relating to engaging digital publics, including through crowdsourcing. Some online biographical projects are being conceived of as virtual research environments in their own right, allowing for sophisticated faceted searching, relationship mapping, analysis and visualisation of results in addition to standard information discovery. These transformations will be considered in the broader context of digital life writing as an emerging field in digital humanities.

Particular reference will be made to the Australian Dictionary of Biography (ADB) online and associated projects, which support innovation in life writing in multiple media formats. Eighteen volumes of the dictionary and one supplementary volume have been published to date, consisting of a total of around 12,500 individual biographical articles. The ADB is the premier reference resource for the study of the lives of Australians who were significant in Australian history, and the largest ever collaborative project in the social sciences and humanities in Australia.

ODDly Pragmatic: Documenting Encoding Practices in Digital Humanities Projects

James Cummings (University of Oxford)

Use of the TEI Guidelines for Electronic Text Encoding and Interchange is often held up as the gold standard for Digital Humanities textual projects. These Guidelines describe a wide variety of methods for encoding digital text and in some cases there are multiple options for marking up the same kinds of thing. The TEI takes a generalistic approach to describing textual phenomena consistently across texts of different times, places, languages, genres, cultures, and physical manifestations, but it simultaneously recognises that there are distinct use cases or divergent theoretical traditions which sometimes necessitate fundamentally different underlying data models. Unlike most standards, however, the TEI Guidelines are not a fixed entity as they give projects the ability to customise their use of the TEI -- to constrain it by limiting the options available or extending it into areas the TEI has not yet dealt with. This combination of the generalistic nature and ability to customise the TEI Guidelines is both one of its greatest strengths as well as one of its greatest weaknesses: it makes it extremely flexible, but it can be a barrier to the seamless interchange of digital text from sources with different encoding practices.

Every project using the TEI is dependent upon some form of customisation (even if it is the 'tei_all' customisation with everything in it that the TEI provides as an example). The TEI method of customisation is written in a TEI format called 'ODD', or 'One Document Does-it-all', because from this one source we can generate multiple outputs such as schemas, localised encoding documentation, and internationalised reference pages in different languages. A TEI ODD file is a method of documenting a project's variance from any particular release of the full TEI Guidelines. This same format underlies the TEI's own steps towards internationalisation of the TEI Guidelines into a variety of languages (including Japanese). The concept of customisation originates from a fundamental difference between the TEI and other standards -- it tries not to tell users that if they want to be good TEI citizens they must do something this one way and only that way, but while making recommendations it gives projects a framework by which they can do whatever it is that they need to do but document it in a (machine-processable) form that the TEI understands. Such documentation of variance of practice and encoding methods enables real, though necessarily mediated, interchange between complicated textual resources. Moreover, over time a collection of these meta-schema documentation files helps to record the changing assumptions and concerns of digital humanities projects.

While introducing the concepts outlined above, this paper showcases projects from the University of Oxford as examples of the way in which this customisation framework can be used for real and pragmatic project benefits.

[Key Words: Text Encoding Initiative, XML, Standards, Customization, Schemas]

Designing the Chronological Reference Model to Define Temporal Attributes

Yoshiaki Murao (Nara University), and ***Yoichi Seino*** (Kyoto University)

The standard temporal attributes we can use in information systems are only"year", "month" and "day" in the Gregorian calendar, except shorter than a day. However, we usually categorize the historical issue that is already passed certain years into an "era", "age" or "period." Especially, in the history or archaeology, there are not so much information about the year of historical issues that they belong to and we can only know the "period" of the large majority of them. In current, the definition of era, age or period in the information system is not standardized practically. As it depends on the implementation of applications for each, it may cause the incompatibilities between databases stored and accumulated at many places, in the near future. In this paper, we introduce the "Chronological Reference model" (CR model) which expresses a type of temporal reference system commonly usable to define the era, age, period, generation or other types of temporal ranges and is able to apply them to temporal attributes of every objects in information systems.

There are currently two major international standards for temporal attributes. One is ISO 8601 "Data elements and interchange formats – Information interchange – Representation of Dates and Times." It defines the basis of current rules of day or time in information system. The other is ISO 190108 "Geographic information – Temporal schema." It defines the schema to accept many types of calendars or eras. It also defines ordinal era to support the Jurassic period or the Cretaceous period, that are classified the order of periods but cannot be defined the start or end year of each period. We studied to apply the ordinal reference system of ISO 19108 to archaeological features or finds, and reached to design the CR model.

The chronology is the term that defines and orders some periods. And its characteristics are similar to the ordinal reference system defined by ISO 19108. Hence we define the classes composed an original temporal reference system which inherit some classes as the ordinal reference system model and cover the chronological specifications. New classes defined in the CR model are ChronologicalSet, ChronologicalReferenceSystem, ChronologicalEra and some datatype classes. ChronologicalReferenceSystem class is the subtype of TM_OrdinalReferenceSystem class which is defined in ISO 19108, and ChronologicalEra class is from TM_OrdinalEra class. These classes have additional attributes or relationships to cover the chronological requirements. So the CR model is conformed to the international standard.

We derived the CR model from the requirements of history or archaeology. But it is applicable to many other fields. The information which is aggregated and managed by the computer system will increase enormously in future. So the temporal category will become important to classify them, as same as the general or geographical category. With regard to the "period" which is used in general, we need the common specification which is unified controlled chronological definition and referenced each other. The CR model can be positioned there, and will play an important role when the variety of data is collected with chronological temporal attributes, referred by many researchers and analyzed them from chronological relationships. Actually it is already required now, because the information accumulation has been starting. The CR Model will contribute to the standardization of chronologies to set temporal attribute values appropriately for all information.

[Key Words: Chronology, Temporal schema, Archaeology, History, Geographic Information System (GIS)]

Ideal Type Modelling and Analysis
– A Model Driven Approach in Cultural Sciences –

Yu Fujimoto (Nara University)

In the present day, nations are strongly connected and the importance of international cooperation is highly stressed in global society. However, such cooperation is still not easy despite various efforts to establish a healthy international society. To reach the next stage of a global society, understanding international conditions objectively will be essential. The problem is that each nation (or individual) has their own opinions, which have been derived from different materials, cultural and religious beliefs, and methodologies. Social conditions and these different viewpoints have caused conicts between nations, cultural regions, and/or individuals. Existing historical approaches of "documentation" have not thus far been able to overcome the problem.

"Ideal Type Modelling and Analysis(ITMA) [1]", a methodology proposed by Fujimoto, and based on Max Weber's methodology of "Ideal Types [2]" and Object-oriented modelling (OOM), might offer a solution. The methodology offers a Model Driven Approach (MDA), enabling automated cultural studies. The researcher designs his or her own ideas or cognition to certain phenomena from the aspects of structures and various kinds of behaviours (or operations), and then, his or her aimed database schemas and programmatic source codes both for constructing the database and for analytical functions are generated automatically. Finally, the researcher receives his or her individual research database and the result of the analyses. The first information model designed is the "Ideal Types Model (ITM)", in which the models are stored in a model management repository when the database is constructed. Although the ITMA work-ow works for individual researchers, differences among researchers' ideas would be mathematically analysed by using stored ITM. Although the methodology is still developing and not get at a practical stage, it would provide a modernised-classical methodology for large volume of data in the present age.

In this paper, the author will discuss one of the technical problems for generating programmatic source codes from ITM: how to generate such completed source codes. As with the original idea of Weber's ideal types, ITM should always be changing. Therefore, a method generating runnable programmes seamlessly from the models is required. The author is developing a test module generating the completed, not skeletal, source codes via the XMI (XML Metadata Interchange), which is commonly used in UML interchange. Although the UML/XMI do not specify to fact codes, in the current project, "note" or "annotation" fields for each class are utilised relating runnable source codes. In the annotation area, linkages to Python scripts are indicated, and the completed source codes are encoded when the corresponding XMI is parsed.

The developing methodology of ITMA offers MDA in cultural sciences, and enables objective observation in social, historical and geopolitical phenomena. It provides reproducible and comparable interpretations for others. In this paper, the author discussed how the researchers' cognition is modelled, encoded and generated.

The test module can generate runnable codes from a model. However, strict specifications for denoting linkages to the source codes are required, without changing the standardised grammars of UML and XMI. Additionally, version control methods will also be required if cultural science needs to trace the temporal transformation of each individual researcher's cognition.

[Key Words: Ideal Types Modeling and Analysis (ITMA), Model Driven Research (MDR)]

References

[1] Y. Fujimoto. *Basic Studies for Culture and Information Science & Ideal Type Modelling and Analysis-Concepts and Methods in Culture and Information Science-*. PhD thesis, Department of Culture and Information Science, Doshisha University, Tatara Miyakodani 1-3, Kyotanabe city, Kyoto, Japan, 2010.

[2] M. Weber. *The "Objectivity" of Sociological and Socio-political Knowledge*. 1904.

Exploring the Future of Publishing through Software Prototyping: Three Reports on a Workflow-Management User Experience Study

Geoff G. Roeder, ***Teresa Dobson***, ***Ernesto Peña***, ***Susan Brown***, ***Elena Dergacheva***, ***Ruth Knechtel***, and the INKE Research Group

Large-scale research projects in the digital humanities continue to grow in size and scope as more and more scholars take advantage of the opportunities for collaboration and study afforded by digital communication tools. Such projectsr equire a means to track and sort through large amounts of data, to assign tasks, and to define roles over sometimes large and geographically diverse teams of investigators. The larger the project, the greater is its need forcentralized workflow tracking and management—both for project managers and individual knowledge workers. Moreover, new digital environments have the potential to expand existing practices in the humanities to allow for new forms of engagement with and inter-disciplinary collaboration on both conventional print and born-digital, multimodal content. As McPherson (2013)remarks, these new opportunities challenge us to "… push beyond the idea that there is a single right interface to knowledge or one best way to publish" (p.10). The challenges and opportunities of this new scholarly context form the topic of this panel, explored through three studies of data collected during a recent user experience trial conducted by the Implementing New Knowledge Environments (INKE) group.

As part of INKE's ongoing, multi-year research project into interface design and prototyping (see http://inke.ca/projects/about/), researchers at the University of British Columbia (Vancouver) and University of Alberta (Edmonton) conducted a two-phase study of two closely related project-management software prototypes. Both began as an experiment in the implementation of structured surfaces for visualizing project workflows (Radzikowska et al., 2011). This experimental prototype was customized for twopurposes: 1) to facilitate academic journal editing, and 2) to facilitate the Orlando Project's ongoing use of computing technology to study women's literary history (see http://www.artsrn.ualberta.ca/orlando/). These two medium-fidelity prototypes, *Workflow Editorial Edition* and Workflow Orlando, were then tested independently using the same instruments and experimental design on two separate participant pools (each having relevant professional experience). Participants engaged with the prototypes by following a task list, were encouraged to think aloud as they did so, and then filled out a questionnaire involving both Likert-scale and free response queries. Transcripts of these sessions along with the survey data collected provided information-rich grounds for an analysis of the benefits and challenges for both born-digital projects like Orlando and more conventional academic publishing. They also highlight what practices in interface design and testing can best facilitate users' feeling comfortable with and gaining the benefit of such workflows.

This panel consists of three papers. The first paper, "Collaboration by Design: Institutional Innovation through Interface Aesthetics," discusses the potential for new modes of academic publishing and collaboration afforded by online workflow tools. The second, "Why is it doing that? Structural and ontological interface metaphors in Workflow," discusses observed patterns of behaviour among the participants in relation to theories of interface metaphor and design. The third paper, "Prospects and pitfalls of workflow management in born-digital projects," reports on the experience of expert users from the Orlando Project using the Workflow

Orlando Edition.

[Key Words: user testing; prototype; publishing; workflow; interface design]

‘Collaboration by Design: Institutional Innovation through Interface Aesthetics’

Presenter: ***Geoff G. Roeder*** (University of British Columbia)

Recent experiments in developing scalable, open-source publishing and research support software for the humanities have begun to realize some of the opportunities of the ongoing digital revolution. As Project 5 of the International Association for Visual Culture (IAVC) remarks, recent developments in online publishing present “opportunities and challenges … perhaps the most radical since the development of moveable type and its consequent market in reading” (2013, p.1). The proposed paper reports on the key findings from the *Workflow Editorial Edition* user experience study, with an emphasis on new opportunities for collaboration encouraged by the interface. *Workflow Editorial Edition* is a prototypical browsing interface designed to enable tracking of the scholarly editing process in the context of journal and monograph publication (see Figure 1). The majority of the nine participants were experienced editors and publishers of academic manuscripts and articles. After structured engagement with the project management tool, a few of them imagined new possibilities for collaboration and closer engagement with other members of a publishing team. One participant imagined a wiki-style circular pane in a central visual field to accommodate open and anonymous collaboration on articles rather than the conventional (and frustrating in the participant’s view) submission-revision-resubmission cycle. This highlights how the aesthetics of an interface can simultaneously facilitate collaboration and instantiate a particular model of authorship as an organizational concept, a model summarized by Mandell (2010) as not “author as owner but author as inventor – copyright here would have no place; issuing patents for literary works would be more appropriate” (p. 126). Another participant imagined

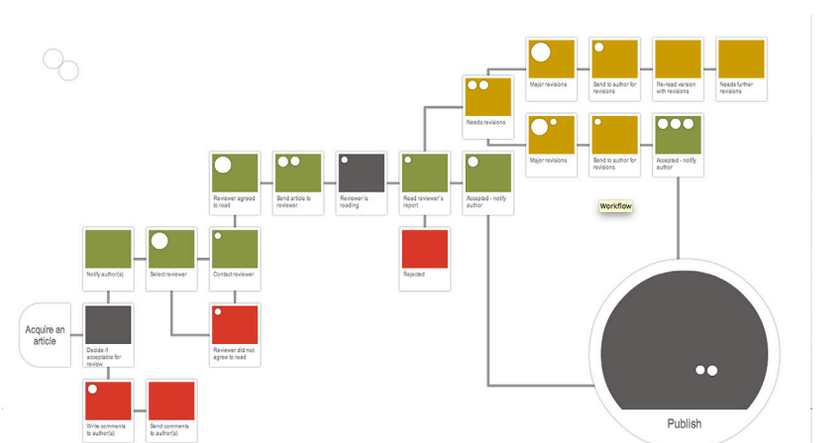


Figure 1: Workflow Editorial Edition interface

an interface that automatically sent ready-made requests and reminders to help a busy editorial team manage the complex professional and interpersonal demands of academic publication. As Fitzpatrick (2013) reminds us, "our communication systems have come to be the way that they are through a series of historical processes and human decisions, and thus … they can be changed" (p.20). *Workflow Editorial Edition* as an interface prototype and as an aesthetic "crafted experience" of a project's progress towards completion can contribute to the Digital Humanities' reimagining of the future of publishing as intentional, designed features of a new generation of academic publishing tools.

'Why is it doing that? Usability and Interface Metaphors in *Workflow*'

Presenter: ***Ernesto Peña*** (University of British Columbia)

Metaphors are a widely used resource for interface design and analysis. Based on Lakoff and Johnson's seminal work on metaphor, Barr (2002) developed a model that acknowledges three types of metaphors commonly used by designers to give individuals who interact with an interface a sense of its logic from first sight, and to scaffold their understanding of it and its general affordances for action. These are known as orientation, ontological, and structural metaphors. As an addition to Lakoff and Johnson's taxonomy of metaphors, Barr proposed two supplementary metaphors that derive from the structural, which he called *element* and *process* metaphors. Generally speaking, it is possible to assert that the first group of metaphors assists a user in becoming acquainted with the function of the interface, and the second group, added by Barr, relates to the actual operation of that interface.

During the analysis of the collected data, it was found that those participants who took what was termed a "reflexive" stance towards the Workflow interface tended to assess the appropriateness of the orientation, ontological, or structural metaphors only. Those participants who decided to take what was termed an "active" stance tended to address issues related to element and process metaphors of the interface instead.

In the current INKE prototype design cycle, lower-fidelity digital prototypes have been favoured over low-fidelity paper or *pdf*-based prototypes as an experiment in interface design and testing. The findings of this study suggest that prototype testing for digital humanities experimental software could productively include one more step in the interface development protocol, prior to the digital prototype stage. The purpose of this phase would be the exploration, proposition, or assessment of possible metaphors at a conceptual level with actual user groups working on low-fidelity representations of the interface. In this stage, based on user group expertise of the editorial process, the participants might participate in focusing the design team towards which resources for meaning-transference are more suitable for the task. After this design process, a digital prototype would then be created to test the congruency between the structural metaphors initially chosen and the actual operation of the interface (e.g.,De Souza et al., 2001,Nadin, 2001, Reilly et al., 2005). Our findings suggest thatsuch approaches would be valuable for use in humanities-orientedsoftware prototyping as well.

‘Prospects and Pitfalls of Workflow Management in Born-digital Projects’

Presenter: ***Elena Dergacheva*** (University of Alberta)

Orlando Workflow is a prototypical browsing interface designed for the Orlando Project, an electronic text base on women’s writing in the British Isles, to enable tracking of the writing and XML encoding–by multiple team members–of interpretive documents (see Figure 2). The user study with Orlando Workflow is the second phase of a two-phase study of workflow management in digital publishing (results from the first phase are addressed in the preceding paper proposals). The eight participants in this phase were primarily Orlando project members who also had substantial scholarly publishing and editing experience. Participants followed a task list in interacting with the interface, offering talk-aloud commentary, and subsequently completed a short questionnaire, consisting of Likert-scale and free response items, about their experience.

Results of the study indicate that digital editing processes may lead to new modes of organizational operation, especially in digital projects with a large scope like Orlando. Key characteristics of this mode include but are not limited to dynamic and changing environments that break the conventionally rigid hierarchies of editing steps, providing for flexible sequence and development. Other features include transparency, speed, and new approaches to collaboration. Additional points of discussion among participants included the following: 1) ethics and transparency of the editorial process; 2) methods of sharing authority between coders and editors as well as among team members in general; 3) decision-making and balance between speed and quality. This paper will review these findings with a view to contemplating the affordances of Orlando Workflow as well as the way in which such flexible approaches to editing may be modifying publication practices in the Digital Humanities with respect to born-digital publications.

Figure 2: Orlando Workflow interface

References

Barr, P., Biddle, R., & Noble, J. (2002). A taxonomy of user-interface metaphors (pp. 25–30). Presented at the Proceedings of the SIGCHI-NZ Symposium on Computer-Human Interaction, ACM.

De Souza, C. S., Barbosa, S. D. J., & Prates, R. O. (2001). A semiotic engineering approach to user interface design. *Knowledge-Based Systems*, 14(8), 461–

465. doi:10.1016/S0950-7051(01)00136-8

Nadin, Mihai. (2001). One cannot not interact. *Semiotic Approaches to User Interface Design*, 14(8), 437–440. doi:10.1016/S0950-7051(01)00138-1

Fitzpatrick, Kathleen. (2013). Scholarly Communication and Scholarly Societies. *Future Publishing: Visual culture in the age of possibility*. Project 5 of the International Association for Visual Culture. Date of access: 5 May 2013. <http://iavc.org.uk/2013/future-publishing-visual-culture-in-the-age-of-possibility>.

Mandell, Laura. (2010). Special Issue: 'Scholarly Editing in the Twenty-First Century' – A Conclusion. *Literature Compass* 7/2 (2010): 120–133.

McPherson, Tara. (2013). Some Theses on the Future of Humanities Publishing, Scholarly and Otherwise. *Future Publishing: Visual culture in the age of possibility*. Project 5 of the International Association for Visual Culture.Date of access: 5 May 2013. <http://iavc.org.uk/2013/future-publishing-visual-culture-in-the-age-of-possibility>.

Project 5. (2013). Introduction. *Future Publishing: Visual culture in the age of possibility*. Project 5 of the International Association for Visual Culture. Date of access: 5 May 2013. <http://iavc.org.uk/2013/future-publishing-visual-culture-in-the-age-of-possibility>.

Reilly, D., Dearman, D., Welsman-Dinelle, M., & Inkpen, K. (2005). Evaluating early prototypes in context: trade-offs, challenges, and successes. *IEEE Pervasive Computing*, 4(4), 42–50. doi:10.1109/MPRV.2005.76

Transcending Borders through DH Networking in the Asia-Pacific

Paul Arthur (University of Western Sydney), ***Jieh Hsiang*** (National Taiwan University), and ***Masahiro Shimoda*** (University of Tokyo)

With the rapid dissemination of collaboration between the humanities and the information technologies in the name of Digital Humanities (DH), many DH-focused research centers and academic associations have been established across the globe. The Asia-Pacific is one of the most fast growing areas in the movement.

There are several centers that are not only working on the cultural and historical contents originated within their own regions, but also contributing to the world-wide DH community by providing e-Research platforms or introducing multimedia computing for the humanities. For example, the Digital Humanities Center for Japanese Arts and Cultures (DH-JAC) at Ritsumeikan University, which started in 2007, emphasizes on conducting synthetic research in DH concerning tangible and intangible cultural heritages. The Research Center for Digital Humanities at National Taiwan University (NTU) was launched in 2008, which has been conducting research projects to digitally preserve important, unique and delicate cultural artifacts and historical resources for promoting digital information exchange globally. Besides these centers, many research institutions and research groups in the area are actively taking responsibilities and initiatives for sharing digital methodsworld-wide, resources and research tools to sustain DH researchers and projects by collaborating with the Alliance of Digital Humanities Organizations (ADHO) and centerNet.

In terms of academic associations, the Australasian Association for Digital Humanities (aaDH) was formed in 2011 to serve a growing digital research community in Australia, New Zealand and more widely in the regions of Australasia and the Pacific. The Japanese Association for Digital Humanities (JADH) was also launched in 2011, which has been aiming to align DH-related domestic associations and promote international collaborative works. The Research Center for Digital Humanities at NTU has been organizing international conferences on Digital Archives and Digital Humanities since 2009 to encourage collaboration between the information technologies and domain experts in the humanities and social studies for accumulating massive digitized collection and promoting interdisciplinary connections.

This panel session invites three leading scholars who are actively involved in fostering and developing DH community in the Asia-Pacific area. They will share the information about the trend in DH-related scholarly activities in each county/region. They will also discuss how collaborative activities and scholarly networking in the area will be able to transcend the particularity of each region for contributing to the knowledge creation in the global DH community.

An Icv nvestigation on Twitter Use of Researchers: User Type Classification and Text Analysis

Yui Arakawa, ***Ryosuke Yoshimoto***, ***Fuyuki Yoshikane*** (University of Tsukuba), and ***Takafumi Suzuki*** (Toyo University)

Gathering information from social media content is becoming increasingly popular. Twitter, a microblog where posts are limited to 140 characters, is an excellent platform for instant and interactive information gathering. Many studies have focused on the role of Twitter in propagating information and spreading rumors (Miyabe et al., 2013; Yamamoto et al., 2012). Other studies have attempted to differentiate between tweets on the basis of genders, number of retweets, and other latent attributes (Arakawa et al., 2012; Burger et al., 2011; Rao et al., 2010). However, there has been less emphasis on the analysis of Twitter posts to obtain information specialized to specific domains. Such analysis could enable simple and rapid identification of information related to state-of-the-art technology.

This study reports preliminary analysis of Twitter posts to obtain domain-specific information. We examine the Twitter user profiles of specialists such as academic researchers and investigate the textual characteristics of their "tweets." We first downloaded 11,900 randomly sampled tweets and 11,577 tweets that were posted by specialists using the Twitter API from July 3, 2012 to August 19, 2012. We manually examined the profiles and applied MeCab (code.google.com/p/mecab) for morphological analysis. We applied random forests machine learning (Breiman, 2001) to classify the randomly sampled tweets and those by specialists, and to extract the distinctive features that influence the classification. We used bag-of-words, number of characters, number of replies (@), number of retweets (RT and QT), number of hash tags (#), number of URLs, and relationship with respondents (whether the respondents were their followers or not) as the features. We evaluated the results of the classification experiments using precision, recall rates, and F1 values (Tokunaga, 1999). We also evaluated the importance of the variables by random forests (Breiman, 2001).

The results of the profile analysis show that most academic researchers use their real names on their profiles and that their positions and affiliations are diverse. The results of random forests classification experiments show that we obtain more than 95% of precision, recall rates, and F1 values. They also show that citation marks (「」), relationship with respondents, and functional expressions such as "toiu" and "nitsuite" are some of the important features. These results indicate that tweets posted by specialists have characteristics that distinguish their tweets from those of others. With qualitative analysis, we conclude that many specialists tweet about their individual activities, education, or researches, and their tweets contain domain-specific knowledge and have identifiable textual characteristics. This study provides basic findings that can be applied to obtain domain-specific knowledge from Twitter. In future, we will explore development of an automated knowledge extraction system for Twitter posts focusing on a specific domain.

[Key Words: computational stylistics, research activities, text analysis, text mining, Twitter]

References

Arakawa, Y. et al. (2012) Stylistic analysis of tweets that are likely to be shared, *JADH2012 Conference Abstracts*, 48.

Breiman, L. (2001) Random forests, *Machine Learning*, 45(1), 5–23.
Burger, J. D. et al. (2011) Discriminating gender on twitter, *Proceedings of EMNLP2011*, 1301-1309.
Tokunaga, T. (1999) *Jouhou Kensaku to Gengo Shori*, University of Tokyo Press, Tokyo.
Miyabe, M. et al. (2013) Characteristic analysis of rumor and correction texts on microblog, *IPSJ Journal*, 54(1), 223-236.
Rao, D. et al. (2010) Classifying latent user attributes in twitter, *Proceedings of SMUC2010*, 37-44.
Yamamoto, M. (2012) Information propagation network for 2012 Tohoku earthquake and tsunami on twitter, Joho Syori, 53(1), 1184-1191.

Validation of Hierarchical Rhetorical Structure by Combination of Quantitative Evaluation and Traditional Rhetorical Analysis

Hajime Murai (Tokyo Institute of Technology)

In order to interpret the Bible precisely, literary criticism is a promising field. It is a methodology for analyzing the Bible as literature and its use of literary techniques. A marked literary characteristic of the Bible is its sophisticated structures, which comprise classic rhetorical structures such as chiasmus, concentric structures, and parallelisms.

There are several merits to identifying rhetorical structures in the Bible. It can clarify the divisions in a text; moreover, correspondences of phrases in them signify deeper interpretation. If the rhetorical structure is the concentric structure, the main theme of that text is also clarified.

However, there are some problems regarding rhetorical structures.

First, there is no clear definition regarding what is a valid correspondence. Some structures correspond by words or phrases, but more abstract themes may also be the element of correspondences. The length of the text unit is not uniform. Some structures are composed of phrases while other structures are composed of pericopes (periscope is a unit of short story in the Bible). Therefore, a quantitative validation method for the rhetorical structure of the Bible is necessary.

In this research, two approaches are combined to validate a hypothesis of a hierarchical rhetorical structure of the Bible. One approach is quantitative evaluation of contingency of a hypothesis. In quantitative evaluation approach, at first the relationships between each part of the text in the rhetorical structure are validated on the basis of the common occurrence of rare words and phrases. At second, the probability of accidental occurrences of common words and phrases in the test hypothesis is calculated. In this approach, two types of random patters are evaluated. One is random combination of pericopes, and the other is random division of pericopes. Three types of random patterns (random combination, random division, and both random) are compared with the test hypothesis and statistical significances are calculated. As a result, rhetorical structures of all long books of the Bible are statistically validated in several types of random patterns.

The second approach is traditional rhetorical analysis by human knowledge. The validity of test hypothesis can be quantitatively calculated by compared with random patterns, but if texts were divided intentionally to make more pairs of rare words and phrases, it may be validated wrongly. Therefore the validity of text division should be examined by some other methodology. In the test hypothesis of hierarchical rhetorical structure, rhetorical structures of whole text size are composed of pericopes. Also those pericopes are basically composed of some type of rhetorical structures (chiasmus, parallelisms, and concentric structure). Then validity of division of text is examined by analysis of pericope level rhetorical structures.

As a result, divisions of pericopes in test hypothesis were validated because some rhetorical structure or some type of rhetorical unit was found in basically all hypothetical pericope.

Though the result of this research is eclectic, those extracted rhetorical structures within pericopes might be quantitatively evaluated in the next step.

[Key Words: Rhetorical Structure, Bible, Chiasm, Quantitative Analysis]

Workset Creation for Scholarly Analysis: Preliminary Research at the HathiTrust Research Center

J. Stephen Downie (University of Illinois at Urbana-Champaign),
Tim Cole (University of Illinois at Urbana-Champaign), ***Beth Plale*** (Indiana University), and
John Unsworth (Brandeis University)

Scholars rely on library collections to support their scholarship. Out of these collections, scholars select, organize, and refine the worksets that will answer to their particular research objectives. The requirements for those worksets are becoming increasingly sophisticated and complex, both as humanities scholarship has become more interdisciplinary and as it has become more digital.

The HathiTrust is a repository that centrally collects image and text representations of library holdings digitized by the Google Books project and other mass-digitization efforts. The HathiTrust corpus contains some 10 million volumes comprising over 3 billion pages. The HathiTrust's computational infrastructure is being built to support large-scale manipulation and preservation of these representations, but it organizes them according to catalog records that were created to enable users to find books in a building or to make high-level generalizations about duplicate holdings across libraries, etc. These catalog records were never meant to support the granularity of sorting and selection or works that scholars now expect, much less page-level or chapter-level sorting and selection out of a corpus of billions of pages.

The ability to slice through a massive corpus consisting of many different library collections, and out of that to construct the precise workset required for a particular scholarly investigation, is the "game changing" potential of the HathiTrust; understanding how to do that is a research problem, and one that is keenly of interest to the HathiTrust Research Center (HTRC), since we believe that scholarship begins with the selection of appropriate resources.

Given the unprecedented size and scope of the HathiTrust corpus—in conjunction with the HTRC's unique computational access to copyrighted materials—we are beginning a project that will engage scholars in designing tools for exploration, location, and analytic grouping of materials so they can routinely conduct computational scholarship at scale, based on meaningful worksets.

"Workset Creation for Scholarly Analysis: Prototyping Project" (WCSA) seeks to address three sets of tightly intertwined research questions regarding:

1) enriching the metadata describing the HathiTrust corpus through mining of the resources themselves and leveraging end-user annotations;
2) augmenting string-based metadata with URIs to leverage external services and Linked Open Data to facilitate discovery and the process of organizing HathiTrust resources into collections and worksets; and,
3) formalizing the notion of collections and worksets in the context of the HathiTrust Research Center.

Building upon the model of the Open Annotation Collaboration, the HTRC is engaging with the digital humanities community to begin development of tools and refine requirements and priorities for enriching and augmenting metadata for the HathiTrust corpus. Concurrently, the HTRC is working closely with the Center for Informatics Research in Science and Scholarship (CIRSS) at the University of Illinois to develop and instantiate a set of formal data models that will be used to capture and integrate the outputs of the funded prototyping projects with the larger HathiTrust corpus.

[Key Words: Collection building, Content analysis, Metadata, Linked open data, Formal models]

Geographical Structure of Medical Delivery in Kyoto, Tokugawa Japan: A Historical GIS Analysis on the Distribution of Medical Practitioners

Mamiko Mataza (Ritsumeikan University), ***Akihiro Tsukamoto*** (The University of Tokushima), and ***Tomoki Nakaya*** (Ritsumeikan University)

This study explored the geographical placement of medical practices in Kyoto during the early modern age—the Tokugawa period (early 17th to mid-19th century)—based on spatial and temporal distributions of medical practitioners.

Geographical Information Systems (GIS) were originally ill-suited for analyzing historical data. It was not until the 21st century that Historical GIS research in Japan developed both in terms of methodology and database creation (Japan Council for Historical GIS Research 2012).We applied the Historical GIS approach to medical delivery and health care in past Japanese societies.

Since Kyoto enjoyed the most advanced medicine until the 18th century, large sets of historical material with detailed information of medical practitioners' addresses remained. Data were obtained from various topographies and directories published during that time. Although the importance of such material has been recognized by Japanese historians, previous studies failed to present meaningful results about geographical aspects (for instance, Kyoto Medical Association 1980: Umihara 1999). There is only one map, based on the directory with a preface written in 1843, aboutthe Distribution of Medical Practitioners (Kyoto City 1973).

The methodology of the study is as follows. We built a database of medical practitioners which included names and addresses based on records published between 1685 and 1867. Data were mainly from the "Kyoto topographies database" (http://www.dhjac.net/db1/books/kyofu/index.html). Other sources were various administration documents and directories. Ten resources were selected from the database from different time-periods, with about 20-year intervals, and maps were drawn of the geographical distribution of medical practitioners by specialty categories (e.g. surgery). As addresses in Kyoto were commonly referred to by a combination of north-south and east-west street names, medical practitioners' addresses from the 10 resources were geocoded using the Historical GIS database of Kyoto streets (Kirimura, Tsukamoto, Yano 2009). This is the application of a method linking attribute data (hold x and y co-ordinates) to spatial data (Gregory & Ell 2007).

Results revealed that medical practitioners were not geographically clustered; other occupation groups tended to concentrate in a few specific areas. The type of medical profession was not the determining factor; rather, it was whether they served lords, especially the Imperial Court, that influenced geographic distribution. In the late 17thcentury, many practitioners, serving the Imperial Court and Shogunate, resided near the Imperial Palace. General practitioners serving regular citizens steadily increased thereafter, and came to serve both lords and the public. In the last days of Tokugawa Shogunate, medical practices near the residences of political people rapidly increased; which indicates that at the beginning and end of the Tokugawa period, power impacted on the practices to some degree. Many practitioners, however, tended to provide medical services on major avenues. A researcher notes that practices were opened along prosperous commercial and industrial

area (Moriya 1978). This indicates that medical practitioners took account of the geographical locations of their medical practice as an important aspect in their management strategy. Importantly, our results suggest that medical practitioners of good reputation were allowed to decentralize through the city and lords did not monopolize them in the early modern age of Kyoto.

[Key Words: GIS, Archive database, Kyoto, early modern, Medical practitioner]

References

Gregory, I. N., & Ell, P. S. (2007). *Historical GIS: Technologies, Methodologies and Scholarship*. Cambridge, London: Cambridge University Press.

Japan Council for Historical GIS Research. (2012). *Historical GIS Perspectives in Japan*. Tokyo: Bensei Shuppan (Ed.).

Kyoto Medical Association. (1980). *Kyoto no igakushi [The history of medicine in Kyoto]*. Kyoto: Shibunkakushuppan (Ed.).

Kyoto City. (1973). *Kyoto no rekishi [The history of Kyoto]*, 6. Tokyo: Gakugeishorin (Ed.).

Moriya, K. (Ed.). (1978). *Kyouishi no rekishi [The history of doctors in Kyoto]*. Tokyo: Kodansha.

Kirimura, T, Tsukamoto, A, Yano, K. (2009). Building a spatiotemporal database of street names in Kyoto, *IPSJ symposium series*, v.2009, no.16, 331 – 338.

Umihara, R. (1999). Juuhachiseiki ranpouigaku no tenkai to sono syakaitekieikyou [The development and social influence of Dutch medicine in the 18th century]. *Annals of the Society for the History of Western Learning in Japan*, 8, 287–301.

Corrective-Detective Features in the Ongoing *Finnegans Wake* Genetic Research Archive Project

Mikio Fuse (University of the Sacred Heart), and
Asanobu Kitamoto (National Institute of Informatics)

In our *Finnegans Wake* Genetic Research Archive Project (started April 2012) we have adopted the TEI rules on digitizing the different kinds of documents involved in the genesis of the book and, using Perl and XSLT for data manipulation, created a prototype of the archive. Diversity of the documents involved and complexity of the irinter-relationship* have been a big challenge to the digital archiving project in general and to the adoption of TEI standard in particular, but so far so good, as far as the basic matters of data storage and basic document views are concerned. Now we are on the second strand of the project to considerhow to make it a practically and sustainably useful tool for *Finnegans Wake* genetic studies.

A crucial point to remember in creating this archive is that it can never be complete. Once the data is digitally installed, currently available data and their inter-relationship always have a huge potential of instigating a discovery of new related document to be newly included in the archive, or of revising the analysis of the relationship of old and new document data. There is also the more common problem of errors and uncertainty of the available data, which often prove correctible and clarifiable by close analysis of the related document data collectively. That is what makes the "corrective-detective" features an essential consideration of our archive. In the proposed poster presentation we should like to demonstrate a couple of "corrective-detective" features of the archive we have been developing.

The archive prototype is being built with Notebook VI.C.2 as a cornerstone because it involves one of the most complex constellations of *Finnegans Wake* textual genetics. It is an amanuensis copy (and a considerably inaccurate one at that) of those units of Joyce's authorial Notebooks VI.B.34, VI.B.2, VI.D.2 and VI.B.6 which he did not cross out in colour crayons (to mark that he has used them in Notesheets and Manuscripts). Of the four parent notebooks of VI.C.2, Notebook VI.D.2 is not extant but can be restored, at least partially (and imperfectly), through analysis of VI.C.2, because, with the considerably inaccurate VI.C.2 entries as clues, we can expect to discover new book or newspaper article Sources from whose words and passages Joyce is supposed to have created the units in the lost Notebook VI.D.2. Once the parent Sources are identified, we can restore the original uncrossed-out units of VI.D.2 by "correcting" the amanuensis' errors in VI.C.2. We can also hope to restore those original VI.D.2 units that were crossed-out (and therefore not copied onto Notebook VI.C.2) by extensive analyses of Joyce's overall patterns of the usage of individual Notebooks in different draft-levels of Notesheets and Manuscripts. By this methodology we have a fair chance of partially restoring both the uncrossed-out and crossed-out unitsof the lost Notebook VI.D.2.

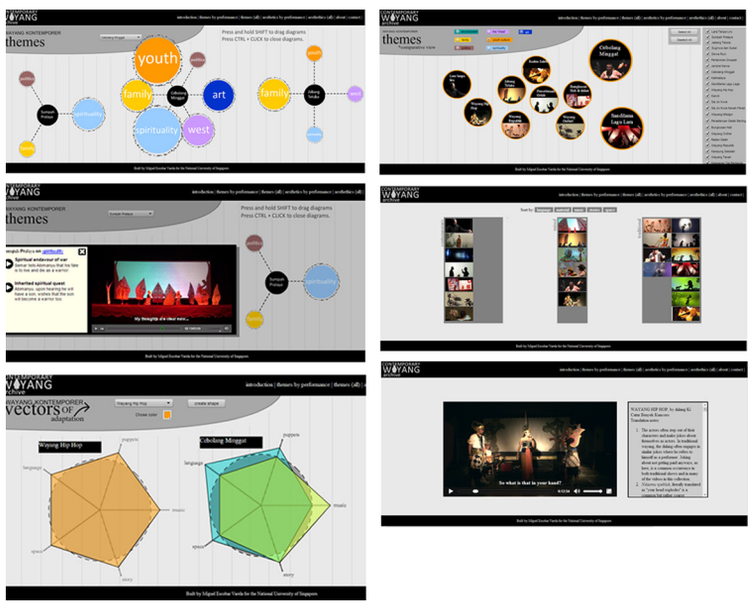
The poster presentation would demonstrate how the archive is designed to positively help the researcher discover new Sources and correct/clarify (or restore) erroneously/uncertainly transcribed (or simply lost) Notebook units.

[Key Words: TEI, digital archive, James Joyce, *Finnegans Wake*, genetic criticism]

*Note: Joyce gathered verbal and conceptual materials for his composition as he readall the miscellaneous books and newspaper articles (Sources) and jotted them down in his Notebooks (B series if extant and D series otherwise). At some stage Joyce also started to ask his amanuensis, Mme. Raphael, to copy the uncrossed-out entries of his authorial Notebooks onto scribal Notebooks (C series). Whether authorial or scribal, the Notebook entries were later transferred to Manuscripts, either directly or via extra-draft Notesheets.

The *Contemporary Wayang Archive*: A Digital Inquiry into the Ethics and Aesthetics of a Theatre Tradition

Miguel Escobar (National University of Singapore)



The *Contemporary Wayang Archive* is a research project aimed at studying a theatre tradition in Indonesia by combining digital research tools, Performance Studies methodologies and an ethnographic approach. The outcome of the research will eventually be presented through a website that includes fully annotated and translated videos of twenty four performances. It also includes different interactive visualizations that allow the users to explore the aggregated data generated from the annotations. These visualizations are linked to time-coded video recordings and to longer ethnographic descriptions. This hybrid presentation style aims at providing both a visual overview of the key research findings and longer, in-depth explanations of the context and specificities of each performance.

These digital tools are particularly fitting to the tradition being studied. Wayang Kulit is the most respected performance tradition in Indonesia, but this research concentrates on radical re-elaborations of this form in the past 15 years. These are analyzed in terms of their *intermedial aesthetics* (the way they use new media in combination with tradition) and of their *post-traditional ethics* (the way they explore changing sociocultural norms).

Despite some similarities, all the performances deploy very different aesthetic strategies and address a variety of issues. This research project aims to compare and contrast these performances and offer an interpretation about what they collectively convey about sociocultural change in contemporary Java. Following a conventional approach in Theatre and Performance Studies, each of the performances was analyzed in terms of their formal qualities and their thematic content, offering explanations of cultural, political and historical references, and developing interpretations of the philosophical and social meaning of the stories.
This analysis was supported by a series of time-coded annotations linked to the videos. The annotations were

then compared and contrasted along different dimensions in order to explore differences and similarities among the performances. Then, the annotations were clustered around several aesthetic dimensions (music, space, actors, language and stories) and ethical themes (such as family, politics and spirituality). In the website that will make the research findings available, the resulting analysis can be accessed in two ways:

- As a series of more conventional essays linked to relevant segments of the videos, where the performance analysis is supplemented by ethnographic descriptions of the ways each performance was received in a Javanese context.
- Through interactive visualizations generated from aggregating the annotations classified under different categories.

Both modes are not mutually exclusive, since the visualizations are linked to the essays and videos. This project required an interdisciplinary approach and a combination of research activities. In order to complete it, I travelled through Java for eighteen months conducting interviews, collecting existing recordings and coordinating new recording efforts. I also edited, translated and annotated the videos and developed a JavaScript interface for the website. In this poster presentation, I will address the implications and possibilities of these mixed strategies, as an instance of DH research into Performance Studies and Anthropological studies of art.

Although the final version of this website is not yet publicly available, a computer with a working version will be available during the poster sessions.

[Key Words: Data Visualization, Video Annotation, Performance Studies, Visual Anthropology]

Automatic Document Layout Analysis based on Machine Learning for Digitizing Japanese Historical Journals

Katsuya Masuda, ***Makoto Tanji***, and ***Hideki Mima*** (University of Tokyo)

This proposal presents an approach of digitizing historical documents especially a process of automatic document layout analysis, logical layout analysis and reading order estimation. Extraction of text from document images is needed for not-digitized historical documents since digitized text data enables more higher-level analysis in digital humanities. For example, new knowledge will be discovered from the results of processing the digitized texts with natural language processing and visualization. Our digitization system receives digital images of documents as inputs, and output the digitized text extracted from the images with logical roles, such as title, author and so on.
The flow of our automated digitization is as follows:
- Recognize the characters in document images with blocks and lines by using a commercially available optical character recognition (OCR) system.
- Estimate logical roles and reading order of blocks in OCR results
- Organizes each article by collecting up the blocks and outputs digitized text with logical roles

The current target resource we deal with is a Japanese journal in humanities, "Shiso (Thought)." The journal has been published since 1921 until now and consists of more than one thousand of numbers, 10,000 articles, 190,000 pages. From the large digitized documents "Shiso," we expect to discover new knowledge of Japanese modern history of philosophy.

In the logical layout analysis step, we proposed a method to identify a logical role of blocks using machine learning technique. We adopted Support Vector Machine (SVM) as a classifier using various types of features of blocks in the OCR results, positions (x and y coordinate), black space length for four directions, block size, character size and linguistic features. The proposed approach were evaluated on the OCR data of "Shiso" documents. The approach classified blocks into five logical roles, "Title," "Author," "Header," "Page Number" and "Text Body," and achieved around 99% F-score in the experiments.

In the reading order estimation step, we introduced Page Splitting method to estimate reading order. The method splits a page vertically or horizontally into two areas which are split recursively until it reaches one block. The different split rule leads to different reading order. To learn the split rule, we used an machine learning optimization method DE (Differential Evolution), and various features, type of split, the number of blocks under a split line, position of a split line, width of a split and logical roles of blocks. Finally, all split blocks are organized using predefined orientation.
Our method estimated reading order with about 0.04 of error (Spearman Distance). The experimental results for the two steps show enough accuracy for digitizing documents.

We constructed a "Shiso" textual database of 80 years from 1921 to 2000 by using our digitizing process, and we also installed it into the knowledge structuring system which is a retrieval system visualizing results in order to help users to discover new knowledge. Although current target is "Shiso," the digitizing process from

digital images of documents can be applied to other targets with low cost to create data with correct logical roles manually for machine learning.

[Key Words: Text Digitizing, Document Layout Analysis, Optical Character Recognition, Machine Learning, Knowledge Structuring]

Lexical Modeling of *Yamabuki* (Japanese Kerria) in Classical Japanese Poetry

Hilofumi Yamamoto (Tokyo Institute of Technology / University of California, San Diego)

This project is a lexical study of classical Japanese poetic vocabulary through network analysis based on graph theory. The analysis is based on co-occurrence patterns, defined as any two words appearing in a poem.

Many scholars of classical Japanese poetry have tried to explain the constructions of poetic vocabulary based on their intuition and experience. As scholars can only demonstrate constructions that they can consciously point out, those that they are unconscious of will never be demonstrated. When we develop a dictionary of poetic vocabulary using only our intuitive knowledge, the description will lack important lexical constructions. In order to conduct more exact and unbiased descriptions, it is necessary to use computer-assisted descriptions of poetic word constructions using co-occurrence weighting methods on corpora of classical Japanese poetry.

We developed the corpora of classical Japanese poetry based on the eight anthologies compiled under imperial order called the "*Hachidaishū*" which were established from ca. 905 to 1205. We also developed a method of co-occurrence weighting (Yamamoto, 2006) which calculates the weight of patterns of any two words appearing in a poem sentence similar to the *tf-idf* method (Sparck Jones, 1972; Robertson, 2004; Manning and Schutze, 1999). The CW allows us to examine the patterns of poetic word constructions through mathematical models.

As a result, when we draw a network model from co-occurrence patterns, we can in general observe a main hub node derived from a topic word. Additionally, we also encounter other hub nodes which do not indicate topic words nor entry items in a poetic dictionary. For instance, when we take *yamabuki* (Japanese kerria) as a topic word and draw its network model, we will observe *kahazu* (frog), *Ide* (place name, proper name), and *yahe* (eightfold or double ower) as hub nodes. The terms *yamabuki*, *kahazu*, and *Ide* are contained in some poetic dictionaries as entry items or collocations. The term *yahe* is, however, not seen in any poetic dictionaries even as a single term. We conclude that a term such as *yahe* can be shown as a hub node which takes an important role to connect a topic word with other peripheral words such as *kokonohe*, *nanahe*, *hitohe*, and plays a supporting role to form a poetic story in the poem even if it is not included in a dictionary.

The finding of this study is that the modeling developed here allows us to 1) discern not only patterns described by experts but also patterns yet undescribed, and 2) identify not only specific or tangible words but also abstract or conceptual words which have a tendency to be left out of dictionaries.

[Key Words: corpus linguistics, co-occurrence weight, visualization, Japanese literature, network modeling]

References

Manning, Christopher D. and Hinrich Schutze (1999) *Foundation of statistical natural language processing*, Cambridge, Massachusetts: The MIT press.

Robertson, Stephen (2004) "Understanding inverse document frequency: on theoretical arguments for IDF", *Journal of Documentation*, Vol. 60, pp. 503-520.

Sparck Jones, Karen (1972) "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", *Journal of Documentation*, Vol. 28, pp. 11{21.

Yamamoto, Hilofumi (2006) "Konpyūta niyoru utamakura no bunseki / A Computer Analysis of Place Names in Classical Japanese Poetry", in *Atti del Terzo Convegno di Linguistica e Didattica Della Lingua Giapponese, Roma 2005* : Associazione Italiana Didattica Lingua Giapponese (AIDLG), pp. 373-382.

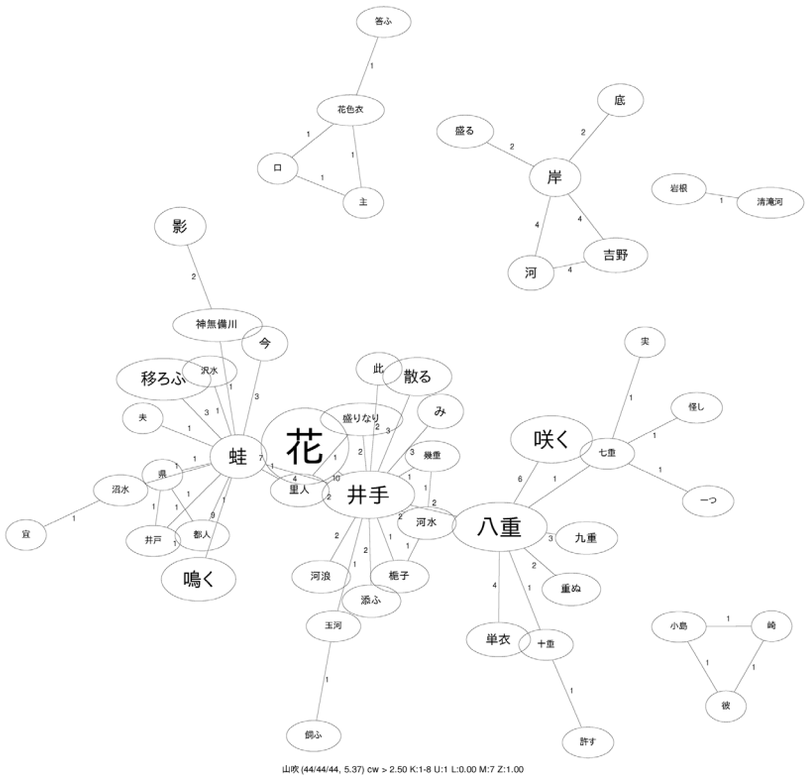


Figure 1: Graph model of *Yamabuki*: a core node, 山吹 *yamabuki*, is pruned. *kahazu* (蛙 , frog), *Ide* (井手 , place name, proper name), and *yahe* (八重 , eightfold or double ower) are observed as hub nodes.

Visualization of the Constitution of Written Language on the Web

Makiko Harada, and ***Hidenori Watanave*** (Tokyo Metropolitan University)

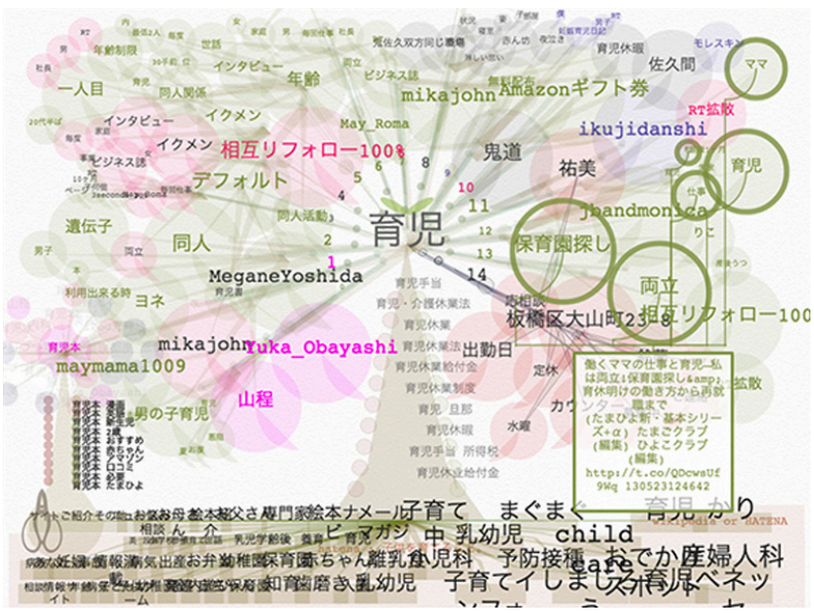


Fig 1:Tree of Language, or "Kotobanoki"

The text on the Web consists of written languages, which reflect the emotions, lifestyles and environments of people from every walk of life. It is organized very differently from a dictionary. Studies have utilized network models to visualize relationships between words on the web. But we should point out that these studies have not paid due attention to the user's emotional state of mind and thus it is difficult to see in an intuitive way the background that gave rise to the context. We thought that the emotional context in the sentence is important to understand the real usage of the word. Also, these interfaces don't follow average people's usages.

The purpose of this study is to show average people the constitution of written language on the web. We visualize the relationship between words to reflect one's emotions, lifestyles and environments. The results are archived in a way in which you are able to see the constitution of the words in terms of chronology in a correlation chart. In order to achieve our purpose, we have developed Visualization Web contents, or what we call “Kotobanomori” in Japanese. It includes three main categories.

1st: Tree of Language, or "Kotobanoki" (Fig 1)

In this category, users develop a concept tree of a typed word. The concept tree is formed by a correlation chart, which includes the results of four search platforms: Twitter, Wikipedia, Google's search engine, and Yahoo and Google's suggested search results. The extracted keywords from the sentences make up nodes. The results are visualized as an actual tree, which can be related to in an intuitive way. Twitter's results are shown as leaves, Yahoo and Google's suggested results are shown as a trunk, Google's search engine's

results are shown as roots, and Wikipedia's results are shown as the ground. Users can compare the results of the words of each search platforms all at once. The emotional valence, or intrinsic appeal or repugnance, of sentences from Twitter, from positive to negative, are obtained in 3 categories: “like - dislike”, “joy - sadness”, “anger - fear”. A color is assigned and mapped, which is determined by a point scale system and the emotional, psychological impact of the color. So the colors of the leaves are determined by the emotional valence of the words.

2nd: Forest of Words or "Kotobanomori"

This category archives the users' search results of the tree of language. The screenshots of each result are gathered into a forest-like arrangement. The words, colors, and shapes are recreated and users can show the history of the results on a calendar and view other users' searches.

3rd: Root of Words, or "Kotobane"

This category visualizes the relationship between the users' search results of the tree of language in a correlation chart. Users can show the process of how words change and how they relate to each other. Also it shows the development of the word's emotional character. We assume that the word's emotional character reflects social customs and world events.

In order to show the effectiveness of our work, we analyze the web traffic to our site and the users' comments. The average visit duration of the web traffic shows that there was enough time for usage, and the number of visitors had amounted to more than 1,000 in 34 days. Here are a couple examples of users' comments, “I can find new relationships between words that I had not imagined. It really got me thinking.” “The shape of the tree is cute. It is good for brainstorming.”

From user comments and the amount of web traffic, we feel that people find our work useful as a way to show relationships between words on the web.

This work can be accessed by the following URL[1]. It is limited to Japanese only.

[Key Words: information design, digital arts, spoken words, archive, data mining]

1 Kotobanomori : http://kokima.sakura.ne.jp/kotobanomori/kotobanoki.htm

Knowledge Structuring with a Topic Map Based on a Philosopher's Texts and Journal Databases

Yu Inutsuka (University of Tokyo)

To address today's complex social issues, like environmental problems, it is necessary to apply the knowledge of more than one discipline. Besides the research to better organize information through computer programming, there are humanities scholars who cognitively reference and organize many fields, including natural and social sciences. With regards to social issues like environmental problems, scholars' selection of information in accordance with human ethical values is important but this work is difficult to share and be developed by others.

In this research, the author suggests a topic map based on philosopher's texts to share understanding of the theory and to build an information platform for further research. Using this methodology, the work of Augustin Berque has been chosen. Berque was originally a cultural geographer and has developed a theory called mesology to analyze human-environment interactions. He referred to many disciplines of arts and sciences, including philosophy, psychology, paleo-anthropology, ethology, neurosciences, etc, in developing this theory. However, because of its interdisciplinary nature, the framework of mesology is highly complex and difficult to understand. Prior research has only evaluated some particular concepts of mesology but dismissed how those concepts are supported by the reference to knowledge from multiple disciplines. In this study, to apply the organization of knowledge in his theory, first a topic map is build based on his texts. Topic maps have the advantages of visualization of the associations and the contents of topics; they have been used in the archival field and very recently in humanities text analysis. Through text analysis, the important concepts, people, and discipline names referenced by Berque are assigned as topics and connected with other related topics. After the topic map is constructed, the topics of disciplines will be linked to the journal database of that discipline. It will create a network of journal databases based on Berque's philosophical conception of humanity, society, and environment. The map will be available for researchers to evaluate, and they can make corrections in accordance with their own understandings. The network will work as an information platform to find the relevant and latest works produced by other disciplines and to engage in interdisciplinary research.

[Key Words: knowledge structuring, topic map, mesology]

A Diachronic and Synchronic Investigation into the Properties of Mid-Rank Words in Modern Japanese

Bor Hodošček (Tokyo Institute of Technology), and
Hilofumi Yamamoto (Tokyo Institute of Technology / University of California)

The present study focuses on the role of mid-rank words in modern Japanese. Mid-rank words are defined as words having an average TF-IDF (term frequencyinverse document frequency) score. Mid-rank words are often overlooked for words with high TF-IDF scores, which act as reliable topic markers. Words with low TF-IDF scores are in turn seen as functional words and often discarded from analysis. Mid-rank words are thus words that do not lean heavily towards the two extremes of topic and function, but include a mixture of both. As such, their exact grammatical function is elusive and still relatively unknown.

In order to determine the properties of mid-rank words, we analyze midrank words on the synchronic and diachronic axes, based on time-series and register-varied modern Japanese corpora, respectively. Thus, the distributional properties of mid-rank words can broadly be compared to those of high- and low-rank words under various conditions.

Time-series data comprising n-grams sampled from blog posts is used to examine the role of mid-rank words in detecting rumor trends. We use Shewart's control charts method of identifying abnormal variations in time series data on n-grams with average TF-IDF scores. Having identified mid-rank words having abnormal frequency spikes, we use a word list classified according to semantic principles (*bunruigoihyou*) to uncover collocational patterns in time. For example, the frequency of the mid-rank word "America", which is otherwise a common word, was observed to spike around October 2008, which roughly corresponds to the period when the U.S. subprime mortgage crisis started to unfold. By observing the changes in collocations before, during, and after the frequency spike, it is possible to quantify what categories of words lead up to such a spike.

The Balanced Corpus of Contemporary Written Japanese is used to examine the role of mid-rank word collocation networks in the description of register differences. While common methods in corpus linguistics use keywords, which often correspond to words with a high TF-IDF, or function words, which often correspond to words with a low TF-IDF, to classify the register of documents, we focus on the distributional differences of mid-rank words in predicting register. We show that mid-rank words are less sensitive to specific topics or functional word usage, and can explain aspects of variation not discernible with topic or function words alone.

In conclusion, we show that mid-ranked words are crucial for a comprehensive account of any word or collocation, especially in the frame of thesaurus and collocation dictionary construction. We also identify areas for further research on the viability of mid-rank words in diachronic and synchronic studies, such as the need for more fine-grained classification of the mid-rank.

[Key Words: Corpus Linguistics, TF-IDF, diachronic analysis, Synchronic analysis, register]

Case Studies of Archiving Textual Information on Natural Disaster: As a Step for Narrative Visualization

Akinobu Nameda, ***Kosuke Wakabayashi***, ***Takuya Nakatsuma***, ***Tomomi Hatano***, ***Shinya Saito***, ***Mitsuyuki Inaba***, and ***Tatsuya Sato*** (Ritsumeikan University)

There is an emergent need to learn lessons from the experiences on the natural disaster. Considering that the natural disaster occurs, it is also important to convey the lessons to the next generation in order to prepare for the next natural disaster such as the strike of Great earthquake. Although numbers of archiving projects has collected large amount of descriptions, it is not easy to extract meaningful information from the piles of textual information. Thus, the present study explores an effective way to learn lessons from large amount of textual information.

We conducted two case studies that were visualizing textual information of the database on the Great earthquakes. Study 1 visualized the information database of the Great Hanshin-Awaji earthquake of 1995, which was created and published by the government of Japan (Cabinet Office, Government of Japan, 2006). In study 2, we created a visualization of textual information in a questionnaire survey data (Kashima city, 2012) on the Great East Japan earthquake. For the visualizations, we used KACHINA CUBE (KC) system (Saito, Ohno, & Inaba, 2009; Ohno, Saito, & Inaba, 2010) that is a web-based platform allowing us to store, plot and display textual information in three-dimensional space. In the study 1 and study 2, we used the KC system that contained 2D geographical area map and time periods in the three-dimensional space. Segmented textual information from the database was plotted as information fragments in the KC systems.

One of the outcomes of the visualization of the studies was that the users could view the quantitative distribution of segmented textual information in the geographical map and time periods. Comparing the number of information fragments across particular places and time periods, the users could reveal how textual information on the Great earthquake was accumulated and infer the reason of the distribution. In the study1, for example, we found that there were relatively less articles on Nagata area in the earlier period after the Great Hanshin-Awaji earthquake hit, although the area was seriously damaged by the earthquake.

Another outcome of the visualizations was that the information including future view will be contributing to learning lessons from the database. On the one hand, the visualization of the database on the Great Hanshin-Awaji earthquake was gaining the factual information on phenomena (Study 1). On the other hand, the visualization of the database on the Great East Japan earthquake allowed us to find the views for the future as well as the facts in the past (Study 2). Concretely in the visualization of study 2, we could see the link between past experience and future view, such as the demanding voice for making evacuation route known that was from the experience of the difficulty when the Great East Japan earthquake struck. In taking the perspective for extracting meaningful information and learning lessons, the outcome having the information on the future views is more useful to gaining meaningful story. With describing and organizing the experiences that are situated in local place and time, the archive visualizing meaningful story would be contributing to understanding human life.

[Key Words: digital archive, natural disaster, narrative, visualization, learning lessons]

References

Cabinet Office, Government of Japan. (2006). Information database of the Great Hanshin-Awaji earthquake. Retrieved October 5, 2011, from http://www.bousai.go.jp/1info/kyoukun/hanshin_awaji/index.html

Kashima city, Ibaraki, Japan. (2012, March). *Survey report collection on the Great East Japan Earthquake*. Planning division of Kashima city. Retrieved 17 October, 2012, from http://city.kashima.ibaraki.jp/info/detail.php?no=5671

Ohno, S., Saito, S., & Inaba, M. (2010). A platform for mining and visualizing regional collective culture. In T. Ishida (Ed.), *Culture and Computing, LNCS* (Lecture Notes in Computer Science), Volume 6259 (pp.189-199), Berlin: Springer.

Saito, S., Ohno, S., & Inaba, M. (2009, December). A platform for visualizing and sharing collective cultural information. Paper presented at the International Conference Digital Archives and Digital Humanities, Taipei, Taiwan.

Visualizing Proclus' Commentary on Plato's Timaeus with Textual Markup

Hiroto Doi (University of Tsukuba)

Although a large number of studies have been made on digital humanities, few studies have attempted to apply digital humanities to philosophy or religious thought.

The present study concentrates on the method to visualize thought from a viewpoint of philosophy and religious studies. In order to do this, Proclus' *Commentary on Plato's Timaeus* will be focused. It is known as a representative work of Proclus (ca. 412-485), who is one of the greatest neoplatonist in late antiquity and head of the Academy. He comments on Plato's *Timaeus* which had a deep influence on western cosmology and cosmogony not only as a natural philosophy but as a theology. But the *Commentary* has a large volume and complicated description, a new method such as visualization is required to interpret Proclus' thought.

Before visualizing, the first to be needed is the analysis of the text. By using CATMA (Computer Aided Textual Markup and Analysis) which can handle ancient Greek text and can group together inflected Greek wordsby querying with tag (e.g. homoiosis which contains inflection and synonym), collocation of important concepts in the *Commentary* is calculated. The conceptssuch as *homoiosis* (assimilation), *henosis* (unification), *psyche* (soul), *nous* (intellect), *euche* (prayer), are chosen from Plato's and Proclus' religious theme "becoming like god". These words are based on my previous paper because the corpus analysis methods were not suitable for Proclus' philosophical text.

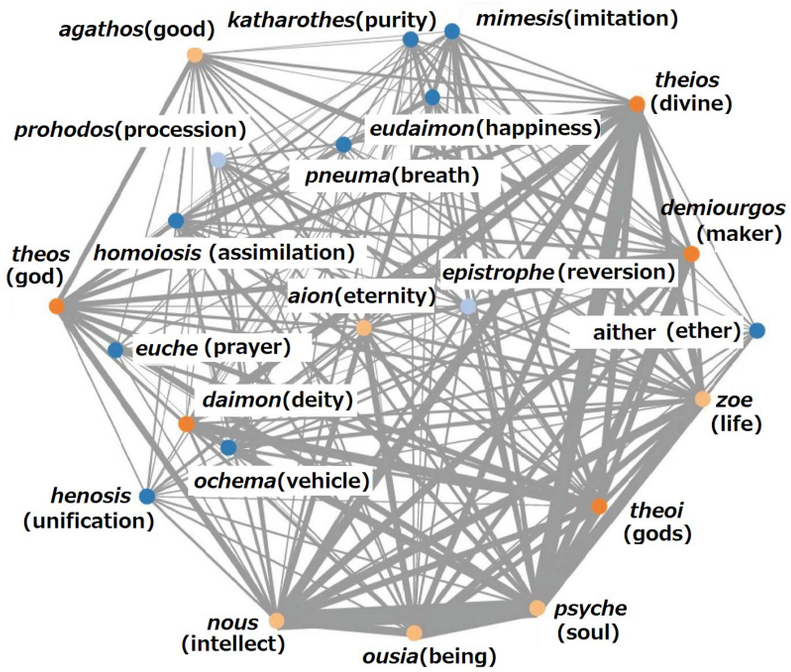
In order to visualize the *Commentary*, this study adopts D3.js (Data-Driven Documents), "a JavaScript library for manipulating documents based on data." The collocation datacalculated by CATMA's query builderare converted to a JSON (JavaScript Object Notation) format file, then the results of visualization is viewed bya web-browser which supports SVG (Scalable Vector Graphics). Among many types of graphs and charts offered by D3.js, this study uses force-directed graph which shows the relationship between nodes (sc. concepts) by lines (stronger line means stronger relationship) and mutual arrangement as below. Though further inquiry into correlation parameters such as Pearson product-moment correlation coefficientby using SPSS was tried, the method for visualization is on the way.

This force-directed graph shows that the concepts concerning "becoming like god" in the *Commentary* connect each other closely and form a complex network, and that religious concepts (e.g.*euche* (prayer), *katharothes* (purity), *henosis* (unification)) are not alienated from important concepts (e.g. ousia (being), *psyche* (soul), *nous* (intellect), *zoe* (life)) in Proclus.

But this visualization also has difficulties. Though the nodes are only 22 in number, this graph is highly complicated to grasp an interrelationship between concepts and the strength of the lines only depends on the number of occurrence, the importance of the connection is not shown.

A further study and sophistication of this visualization of thought should be conducted. But in my poster presentation, this approach will be demonstrated with a laptop computer as a work-in-progress.

[Key Words: visualization, markup, philosophy, religious thought]



Bibliography

Mike Dewar, *Getting Started with D3: Creating Data-Driven Documents*, Sebastopol, O'Reilly, 2012.

E. Diehl (ed.), *Procli diadochi in Platonis Timaevm commentaria*. 3 vols., Leipzig, Teubner, 1903-1906.

E. R. Dodds, *The Elements of Theology: A Revised Text with Translation, Introduction, and Commentary*, Oxford, Clarendon Press, 1963 (Second Edition).

Hiroto Doi, "Spirituality and Religious Thought in Ancient Greek", in: Yoshio Tsuruoka and Hidetaka Fukasawa (eds.), *Spirituality and the History of Religions*, Vol. II, Tokyo, Lithon, 2012.

Manuel Lima, *Visual Complexity: Mapping Patterns of Information*, New York, Princeton Architectural Press, 2011.

Possibilities of the Data Visualization for Humanities in a Web Browser: A Demonstration of the KACHINA CUBE Version.3

Shinya Saito (Ritsumeikan University)

This poster presentation focuses on the KACHINA CUBE version.3 (KC ver.3) system[1], an original Web-based visualization tool. Those who visit the poster presentation can operate the system on the spot.

The first part of the presentation explains the background of the development of KC system. "Visualization" can be divided into two categories[2]. One is “scientific visualization”, and the other is “information visualization”. KC's visualization belongs to the latter. To be more precise, it belongs to its sub-category called "Infographics". In order to develop the methodology of the Inforgraphics further, this research focuses on the following three points. The first point is to automatically reflect the current status of databases, as the CMS (Contents Management System) does. The second is to make the user can operate graphics with interactive visualization. The last point is to use 3-D CG on the browser efficiently.

The next part discusses the purpose of the development. KC's visualization aims to create a sort of “space nature” to a dataset. This can give us new attitude towards various data. Moreover it can support us to construe a large amount of information.

As for the implementation methodology of the KC (version.2), it has an information-viewer, cube geometry to achieve the above-mentioned purpose. Using this, we can put the map that we made into the bottom of cube. This map decides the position of information in two dimensions, and the third dimension decides the time flow of the information. With this KC system, we can store, plot, and display arbitrary information in itsviewer. In the Humanities research, it is essential to sort data out along the timeline, which the KC can visually support with its 3-D CG. Actually, the KC has been helping researching local history and analyzing development of some industries.

Moreover I will explain the difference between KC ver.3 and the earlier version KC ver.3 has some innovative functions for data visualization in a Web browser. In the end, we will discuss how effective the application of KC could be. Our past system operation made two types of merits of KC clear. In other words, KC proves its effectiveness as both a presentation tool and an analysis method.

[Key Words: Visualization, Inforgrafics, CMS (Contents Management System), 3D-CG]

Bibliography

[1] Shinya Saito. “Web Technology and Visual Expression: From a Point of View of e-Research”, In Inaba, M (ed.), *Digital Humanities Research and Web Technology*, Nakanishiya Publishing, 2012, pp.197-214.

[2]Riccard Mazza. *Introduction to Information Visualization*. Springer, 2009.

Situated and Collaborative Learning in 3D Metaverse: A Case Study of Computer-Mediated Cultural Exchange between Japan and Hawaii

Michiru Tamai, ***Mitsuyuki Inaba***, ***Koichi Hosoi***, ***Akinori Nakamura***, ***Masayuki Uemura***, and ***Ruck Thawonmas*** (Ritsumeikan University)

This research introduces a case study of the computer-mediated cultural exchange between Hawaii and Japan on the Internet. It is an outcome of our ongoing research efforts on implementing 3D Metaverse environment for preserving and sharing Japanese traditional culture.

3D Metaverse is a platform for constructing immersive virtual spaces with digitized objects and architectures on the Internet. An avatar, or a controllable virtual character in the space, enables the visitors to the 3D space participating in embodied social interaction with other avatars.

These features of the Metaverse are beneficial to the implementation of learning platforms for those who have difficulties to visit a foreign country or distant area. Therefore, it opens the door to new way of cultural learning.

Also it provides powerful platform for inheriting and sharing traditional tangible and intangible cultural heritage from socio-cultural perspective, which was difficult to provide by conventional web-based e-Learning environments.

First, we demonstrate our Metaverse environment in SecondLife (SL), which is the most popular Metaverse infrastructure. The environment has been constructed based on the results of the questionnaire survey to international students. The space includes Japanese tangible cultural properties such as Shinto shrine, Buddhism temple, and Noh stage. It also has virtual museums for Kimono costume or Yuzen textile design. Various intangible cultural heritages are also digitized and preserved in the space.

Second, we provided a collaborative environment where students will acquire and practice the necessary communication skills under the constructivist principles of situated learning and collaborative learning by the SL. In the experiments, Hawaiian and Japanese students interacted, and exchanged their questions, impressions, or interpretations on traditional culture and custom on each virtual island. We analyzed their interaction and learning process in a virtual world. In the process, the situational cultural learning between newcomers and old timers was enabled.

Finally, we discuss the advantages and limitations of the learning process in our Metaverse platform in terms of situated learning and collaborative learning.

[Key Words: Situated Learning, Collaborative Learning, Metaverse, e-Learning, Japanese Culture]

Spatial Analysis and Web-based Application of "Large-scale Maps of Kyoto City"

Naomi Akaishi (Ritsumeikan University), ***Toshikazu Seto*** (University of Tokyo), ***Yukihiro Fukushima*** (Kyoto Prefectural Library and Archives), and ***Keiji Yano*** (Ritsumeikan University)

This paper analyzes how the City of Kyoto used its land in the Modern period and explores applicability of web maps with an example of digitalized "Large-scale Maps of Kyoto City (*Kyoto-shi meisai-zu*)". Made between 1927 and 1951 for fire insurance purposes, these maps contain various kinds of building-related information necessary to prevent fire.

GIS database of "Large-scale Maps of Kyoto City" is composed of about 160,000 building polygons. What the maps tell us is that, while an average building occupied approximately 93 square meters, about 13% of all the buildings were much smaller, 30 square meters or less each, whose shapes indicated that they must be tenement houses. The maps include about 147,000 buildings with information about their usage and the number of stories on, which are set to work as point data.

The majority of the buildings were one or two stories high, which accounted for more than 80% of the total. High-rise buildings of five stories or more were factories and theatres, which must be regarded as major landmarks during the time period between 1945 and 1955 after WWII. It is worth noticing that descriptions and colors on the maps indicate how buildings were used, most of which were employed for public facilities such as houses, offices, factories, temples and shrines, and schools. Some exceptions include textile-related buildings, pleasure quarters, and mansion houses, which means that the maps functioned as residential ones.

Many of the buildings depicted in "Large-scale Maps of Kyoto City" no longer exist. We publish the Maps' digital images through the ArcGIS online, overlaying them on current housing maps. This service is available not only on the web page but also on iOS or Android dedicated mobile applications downloaded. With this portability, we have started using the maps as a tool to inspire local residents' memories and for social events that involve walking town. In the future, we hope to include more detailed information, such as photographic images of major facilities and descriptions of major events.

[Key Words: "Large-scale Maps of Kyoto City", GIS, digitalization, landscape reconstruction, modern Kyoto]

JADH 2013 & DH-JAC 2013 CONFERENCE ABSTRACTS

Published by the International Institute for Digital Humanities, Tokyo, Japan

ISBN 978-4-9906708-3-2